Cédric Villani

Optimal transport, old and new

September 27, 2006

Springer
Berlin Heidelberg New York
Hong Kong London
Milan Paris Tokyo

Do mo chuisle mo chroí, Aëlle

This is the current version (as of September 27, 2006) of my lecture notes for the 2005 Saint-Flour summer school. There have been important changes from the previous preliminary versions, especially in the last part. A few references and figures are still missing.

This text is much more ambitious than I initially intended it to be, so its publication will be delayed. Presently, everything has been revised except for Chapters 24 and 25 (gradient flows) which still need some reworking; also Chapter 23 needs some more work, and Chapter 22 should be debugged. The bibliography should be completed.

cvillani@umpa.ens-lyon.fr

Contents

Preface

When I was first approached for the 2005 edition of the Saint-Flour Probability Summer School, I was intrigued, flattered and scared.[1] Apart from the challenge posed by the teaching of a rather analytical subject to a probabilistic audience, there was the danger of writing a remake of my recent book *Topics in Optimal Transportation*.

However, I gradually realized that I was offered a unique opportunity to rewrite the whole theory from a different perspective, with alternative proofs and different focus, and a more probabilistic presentation; plus the incorporation of recent progress. Among the most striking of these recent advances, there was the rising awareness that Mather's minimal measures had a lot to do with optimal transport, and that both theories could actually be embedded in a single framework. There was also the discovery that optimal transport could provide a robust synthetic approach to Ricci curvature bounds. These links with dynamical systems on one hand, differential geometry on the other hand, were only briefly alluded to in my first book; here on the contrary they will be at the basis of the presentation. To summarize: more probability, more geometry, and more dynamical systems. Of course there cannot be more of everything, so in some sense there is less analysis and less physics, and also there are fewer digressions.

So these notes are by no means a reduction or an expansion of my book, but should be regarded as a complementary reading. Both sources can be read independently, or together, and hopefully the complementarity of points of view will have pedagogical value.

The text is divided in many short chapters, separated into three main parts. The first part is devoted to a qualitative description of optimal transport; the second part discusses the use of optimal transport in Riemannian geometry; finally the third part is devoted to recent research about a synthetic treatment of Ricci curvature bounds, based on optimal transport. Throughout the book I have tried to optimize the results, the proofs and the presentation, and to provide complete and self-contained proofs of the most important results. Many statements and theorems have been written and proven specifically for this course, and many results appear in rather sharp form for the first time. I also added several Appendices, some of them to present some domains of mathematics to non-experts, some of them to provide proofs of important auxiliary results. All this has resulted in a rapid growth of the document, which in the end is about five times (!) the size that I had planned initially. So the non-expert reader is advised to skip long proofs at first reading, and concentrate on explanations, statements and examples. Also I have tried to present rather comprehensive bibliographical notes, a dauntingly difficult task in view of the rapid expansion of the literature.

About terminology: For some reason I decided to switch from "transportation" to "transport", but this really is a matter of taste.

For people who are already familiar with the theory of optimal transport, here are some more serious changes.

The dynamical point of view is given a prominent role from the beginning, with McCann's concept of displacement interpolation. A recurring idea in these notes is that it is often better to consider optimal transport from some intermediate time, rather than from the initial time. Displacement interpolation is discussed before any theorem about the solvability of the Monge problem, in an abstract setting of "Lagrangian action", which generalizes the notion of length space. This point of view encompasses at the same time recent developments dealing with optimal transport in length spaces, and those about smooth Lagrangian cost functions on Riemannian manifolds. For the convenience of the readers, I included basic reminders about Riemannian geometry and length spaces.

[1] Fans of Tom Waits may have identified this quotation.

I wrote down in detail some important estimates by Mather, well-known in certain circles, and made extensive use of them, in particular to prove the Lipschitz regularity of "intermediate" transport maps (taking as starting measure not the measure at time 0, but the measure at time $t_0 \in (0,1)$). Then the absolute continuity of displacement interpolants comes for free, and this gives a more unified picture of the Mather and Monge-Kantorovich theories. I also rewrote in this way the classical theorems of solvability of the Monge problem for quadratic cost in Euclidean space. Finally, this approach allows to treat changes of variables formulas associated with optimal transport by means of changes of variables that are Lipschitz, and not just with bounded variation.

The links between Ricci curvature and optimal transport, which are discussed at length in Parts II and III, appear for the first time in lecture notes form; many results actually appear for the first time at all. I included reminders about the meaning and calculus with Ricci curvature, and also about the Gromov–Hausdorff convergence; so hopefully the reader can follow the whole discussion without being forced to consult some independent source on differential geometry.

There are several important parts of the theory which I chose not to develop too much, or not at all. In my opinion the topic which is most poorly represented is the regularity theory for optimal transport; first because it is a quite long story, and secondly because it is not necessary for the purpose of these notes. There have been important progress quite recently on the topic of regularity of optimal transport of manifolds (with works by Neil Trudinger, Xu-Jia Wang, Grégoire Loeper and others) and a consistent picture of regularity seems to be emerging right now. A few years from now, an up-to-date monograph on the subject would actually be welcome.

Another topic which is not addressed at all is the numerical simulation of optimal transport. Besides classical methods such as the simplex algorithm, there are more original methods such as the "auction algorithm" by Bertsekas, and more recently numerical methods based on the Monge–Ampère equation. This subject too is in need of a good synthesis work.

Still another subject which is poorly developed is the Monge-Mather-Mañé problem arising in dynamical systems, and including as a variant the optimal transport problem when the cost function is a distance. This topic is addressed in several books on theoretical Lagrangian mechanics, such as Albert Fathi's lecture notes; but now it would be desirable to rewrite everything in a framework that also encompasses the optimal transport problem. An important step in this direction was recently performed by Patrick Bernard and Boris Buffoni. There will be in these notes an introduction to Mather's approach, but there would be much more to say.

The use of optimal transport in the theory of concentration of measure is addressed in Chapter 22, as part of Riemannian geometry — but some of the estimates are established in more general metric structures. About this application of optimal transport one can also consult two recent synthesis documents about concentration of measure: the monograph by Michel Ledoux, *The Concentration of Measure Phenomenon*; and the Saint-Flour course by Pascal Massart.

Gradient flows with respect to the Wasserstein distance are discussed in Chapters 23 to 25. The exposition is, I hope, precise, and addresses the most important points, but it is not exhaustive. In fact, this topic has inflated so quickly in the past five years or so, that an up-to-date and complete reference would fill up a whole book. It was actually the subject of an excellent reference treatise: *Gradient flows in metric spaces and in the space of probability measures*, by Luigi Ambrosio, Nicola Gigli and Giuseppe Savaré. Still in that

reference the emphasis is on Euclidean (or Hilbert) spaces, while now it would be desirable to rewrite the whole theory in a genuinely Riemannian context.

There are other classical applications of optimal transport to various fields of probability theory, which are missing from this book, but can be found in the two-volume treatise by Svetlozar Rachev and Ludger Rüschendorf, *Mass Transportation Problems*.

During the preparation of this text I asked help from a number of people. Among them, Luigi Ambrosio and John Lott are the ones whom I most put to contribution; these notes owe a lot to their detailed comments and suggestions. Most of Part III, but also significant portions of Parts I and II, are made up with ideas taken from my collaborations with John, which started in 2004 as I was enjoying the hospitality of the Miller Institute in Berkeley. Long discussions with Patrick Bernard and Albert Fathi allowed me to get the links between the modern theory of optimal transport and Mather's theory, which were a key to the presentation in Part I. Apart from these people, I received valuable help from François Bolley, Yann Brenier, Dario Cordero-Erausquin, Denis Feyel, Alessio Figalli, Sylvain Gallot, Wilfrid Gangbo, Diogo Gomes, Arnaud Guilin, Michel Ledoux, Grégoire Loeper, Robert McCann, Felix Otto, Ludger Rüschendorf, Giuseppe Savaré, Karl-Theodor Sturm, Anthon Thalmaier, Hermann Thorisson, Süleyman Üstünel, Xu-Jia Wang, and others.

Short versions of this course were tried on mixed audiences in the Universities of Bonn, Dortmund, Grenoble and Orléans, as well as the CIRM in Luminy and the Borel seminar in Leysin. All these institutions are warmly thanked.

It is a pleasure to thank Jean Picard for all his organization work on the 2005 Saint-Flour summer school. Additional thanks are due to the participants for their questions, comments and bug-tracking, in particular Sylvain Arlot (who also did a great job in spotting typos and mistakes), Fabrice Baudoin, Jérôme Demange, Steve Evans (whom I also thank for his beautiful lectures), Christophe Leuridan, Jan Obłój, Erwann Saint-Loubert Bié, and others. I extend these thanks to the joyful group of young PhD students and maîtres de conférences with whom I spent such a good time on excursions, restaurants and other activities, making my stay in Saint-Flour truly wonderful (with special thanks to my personal driver, Stéphane Loisel, and my table tennis sparring-partner, François Simenhaus).

Typing of these notes was entirely performed on my faithful laptop, a gift of the Miller Institute. My eternal gratitude goes to those who made fine typesetting accessible to every mathematician, most importantly Donald Knuth for TEX, and also the developers of LATEX, BIBTEX and `XFig`.

As usual, I encourage all readers to report mistakes and misprints. After publication, **I will maintain a list of errata, accessible from my Web page**.

Cédric Villani
Lyon, September 2006

Conventions

Axioms

I use the classical axioms of set theory; not the full version of the axiom of choice (only the classical version of "countable dependent choice").

Sets and structures

Id is the identity mapping, whatever the space.

If f and g are two functions, then (f, g) is the function $x \longmapsto (f(x), g(x))$. The composition $f \circ g$ will often be denoted $f(g)$.

$\mathbb{N}$ is the set of *positive* integers: $\mathbb{N} = \{1, 2, 3, \dots\}$. A sequence is denoted either $(x_k)_{k\in\mathbb{N}}$, or simply, when no confusion seems possible, (x_k).

$\mathbb{R}$ is the set of real numbers. When I write $\mathbb{R}^n$ it is implicitly assumed that n is a positive integer. The Euclidean scalar product between two vectors a and b in $\mathbb{R}^n$ is denoted indifferently $a \cdot b$ or $\langle a, b\rangle$. The Euclidean norm will be denoted simply $|\cdot|$, independently of the dimension n.

$M_n(\mathbb{R})$ is the space of real $n \times n$ matrices, and I_n the $n \times n$ identity matrix. The trace of a matrix M will be denoted by $\mathrm{tr}\, M$, its determinant by $\det M$, its adjoint by M^*, and its Hilbert-Schmidt norm $\sqrt{\mathrm{tr}\,(M^*M)}$ by $\|M\|_{\mathrm{HS}}$ (or just $\|M\|$).

Unless otherwise stated, Riemannian manifolds appearing in the text are finite-dimensional, smooth and complete. If a Riemannian manifold is given, I shall usually denote by n its dimension, by d the geodesic distance on M, and by vol the volume ($= n$-dimensional Hausdorff) measure on M. The norm on a tangent space of a Riemannian manifold will most of the time be denoted by $|\cdot|$, as in $\mathbb{R}^n$, without explicit mention of the point at which the norm is taken. (The symbol $\|\cdot\|$ will be reserved for special norms or functional norms.)

If Q is a quadratic form defined on $\mathbb{R}^n$, or on the tangent bundle of a manifold, its value on a (tangent) vector v will be denoted by $\big\langle Q \cdot v,\, v\big\rangle$, or simply $Q(v)$.

The open ball of radius r and center x in a metric space $\mathcal{X}$ is denoted indifferently by $B(x, r)$ or $B_r(x)$. If $\mathcal{X}$ is a Riemannian manifold, the distance is of course the geodesic distance. The closed ball will be denoted by $B(x, r])$ or $B_{r]}(x)$. The diameter of a metric space $\mathcal{X}$ will be denoted $\mathrm{diam}\,(\mathcal{X})$.

The closure of a set A in a metric space will be denoted by $\overline{A}$ (this is also the set of all limits of sequences with values in A).

A metric space $\mathcal{X}$ is said to be *locally compact* if every point $x \in \mathcal{X}$ admits a compact neighborhood; and *boundedly compact* if every closed and bounded subset of $\mathcal{X}$ is compact.

A map f between metric spaces $(\mathcal{X}, d)$ and $(\mathcal{X}', d')$ is said to be C-Lipschitz if $d'(f(x), f(y)) \le C\, d(x, y)$ for all x, y in $\mathcal{X}$. The best admissible constant C is then denoted by $\|f\|_{\mathrm{Lip}}$.

A map is said to be locally Lipschitz if it is Lipschitz on bounded sets, *not necessarily compact* (so it makes sense to speak of a locally Lipschitz map defined almost everywhere).

A curve in a space $\mathcal{X}$ is a continuous map defined on a subinterval of $\mathbb{R}$, into $\mathcal{X}$. In these notes the words "curve" and "path" will be synonymous. The time-t evaluation map e_t is defined by $e_t(\gamma) = \gamma_t = \gamma(t)$.

If γ is a curve defined from an interval of $\mathbb{R}$ into a metric space, its length will be denoted $\mathcal{L}(\gamma)$, and its speed by $|\dot\gamma|$; definitions are recalled on p. 89.

Unless otherwise stated, geodesics are *minimizing*, *constant-speed* geodesic curves. If $\mathcal{X}$ is a metric space, the space of all geodesics $\gamma : [0, 1] \to \mathcal{X}$ will be denoted by $\Gamma(\mathcal{X})$.

Being given x_0 and x_1 in a metric space, I denote by $[x_0, x_1]_t$ the set of all t-barycenters of x_0 and x_1, as defined on p. 242. If A_0 and A_1 are two sets, then $[A_0, A_1]_t$ stands for the set of all $[x_0, x_1]_t$ with $(x_0, x_1) \in A_0 \times A_1$.

Function spaces

I denote by $C(\mathcal{X})$ the space of continuous functions $\mathcal{X} \to \mathbb{R}$, by $C_b(\mathcal{X})$ the space of bounded continuous functions $\mathcal{X} \to \mathbb{R}$; and by $C_0(\mathcal{X})$ the space of continuous functions $\mathcal{X} \to \mathbb{R}$ converging to 0 at infinity; all of them are equipped with the norm of uniform convergence. Then $C_b^k(\mathcal{X})$ is the space of k-times continuously differentiable functions $u : \mathcal{X} \to \mathbb{R}$, such that all the partial derivatives of u up to order k are bounded; and it is equipped with the norm given by the supremum of all norms $\|\partial u\|_{C_b}$, where ∂u is a partial derivative of order at most k; $C_c^k(\mathcal{X})$ is the space of k-times continuously differentiable functions with compact support; etc. When the target space Y is not $\mathbb{R}$, the notation is transformed in an obvious way: $C(\mathcal{X}; Y)$, etc.

I use the standard notation L^p for the Lebesgue space of exponent p; the space and the measure will often be implicit, but clear from the context.

Calculus

The derivative of a function $u = u(t)$, defined on an interval of $\mathbb{R}$ and valued in $\mathbb{R}^n$ or in a smooth manifold, will be denoted by u', or more often by $\dot{u}$. If u is a function of several variables, the partial derivative with respect to the variable t will be denoted by $\partial_t u$, or $\partial u / \partial t$. The notation u_t *does not* stand for $\partial_t u$, but for $u(t)$.

The gradient operator will be denoted by grad or simply ∇; the divergence operator by div or $\nabla \cdot$; the Laplace operator by Δ; the Hessian operator by Hess or ∇^2 (so ∇^2 does *not* stand for the Laplace operator). The notation is the same in $\mathbb{R}^n$ or in a Riemannian manifold. Δ is the divergence of the gradient, so it is typically a *nonpositive* operator. The value of the gradient of f at point x will be denoted either $\nabla_x f$ or $\nabla f(x)$. The notation $\widetilde{\nabla}$ stands for the *approximate gradient*, defined in Definition 10.2.

When T is a map $\mathbb{R}^n \to \mathbb{R}^n$, the notation ∇T stands for the Jacobian matrix of T, that is the matrix of partial derivatives $(\partial T_i / \partial x_j)$ $(1 \leq i, j \leq n)$.

All these differential operators will be applied to (smooth) functions but also to measures, by duality. For instance, the Laplacian of a measure μ is defined via the identity $\int \zeta \, d(\Delta \mu) = \int (\Delta \zeta) \, d\mu$ $(\zeta \in C_c^2)$. The notation is consistent in the sense that $\Delta(f \mathrm{vol}) = (\Delta f) \, \mathrm{vol}$. Similarly, I shall take the divergence of a vector-valued measure, etc.

The notation $f = o(g)$ means $f/g \longrightarrow 0$ (in an asymptotic regime that should be clear from the context), while $f = O(g)$ means that f/g is bounded.

The positive and negative parts of $x \in \mathbb{R}$ are defined respectively by $x_+ = \max(x, 0)$ and $x_- = \max(-x, 0)$; both are nonnegative, and $|x| = x_+ + x_-$. The notation $a \wedge b$ will sometimes be used for $\min(a, b)$. All these notions are extended in the usual way to functions and also to signed measures.

Probability measures

The Dirac mass at point x is denoted by δ_x.

The law of a random variable X is denoted by just law (X).

All measures considered in the text are Borel measures on **Polish spaces**, which are complete, separable metric spaces, equipped with their Borel σ-algebra. I shall usually not use the completed σ-algebra, except on some rare occasions in Chapter 5, and this will be emphasized in the context. These issues only arise with discontinuous cost functions, that play a marginal role in these notes; for continuous cost functions the whole theory can be developed with just Borel measurable sets.

A measure is said to be finite if it has finite mass, locally finite if it attributes finite mass to compact sets.

The space of Borel probability measures on $\mathcal{X}$ is denoted $P(\mathcal{X})$, the space of finite Borel measures by $M_+(\mathcal{X})$, the space of finite signed Borel measures by $M(\mathcal{X})$.

The integral of a function f with respect to a probability measure μ will be denoted indifferently $\int f(x)\, d\mu(x)$ or $\int f(x)\, \mu(dx)$, or $\int f\, d\mu$.

If μ is a Borel measure on a topological space $\mathcal{X}$, then a set N is said to be μ-negligible if N is included in a Borel set of zero μ-measure. Then μ is said to be concentrated on a set C if $\mathcal{X} \setminus C$ is negligible. If C itself is Borel measurable, this is of course equivalent to $\mu[\mathcal{X} \setminus C] = 0$.

If μ is a Borel measure, its support $\mathrm{Spt}\, \mu$ is the smallest *closed* set on which it is concentrated. The same notation Spt will be used for the support of a continuous function.

If μ is a Borel probability measure on $\mathcal{X}$, and T is a Borel map $\mathcal{X} \to \mathcal{Y}$, then $T_\# \mu$ stands for the image measure (or push-forward) of μ by T. It is a Borel probability measure on $\mathcal{Y}$, defined by $(T_\# \mu)[A] = \mu[T^{-1}(A)]$. Depending on the authors, the measure $T_\# \mu$ is often denoted by $T\# \mu$, $T_* \mu$, $T(\mu)$, $T\mu$, $\mu \circ T^{-1}$, μT^{-1}, or $\mu[T \in \cdot]$.

The weak topology on $P(\mathcal{X})$ (or topology of weak convergence, or narrow topology) is induced by convergence against $C_b(\mathcal{X})$, i.e. *bounded continuous* test functions. If $\mathcal{X}$ is Polish, then the space $P(\mathcal{X})$ itself is Polish. Unless explicitly stated, I do not use the weak-$*$ topology of measures (induced by $C_0(\mathcal{X})$ or $C_c(\mathcal{X})$).

When a probability measure is clearly specified by the context, it will sometimes be denoted just by $\mathbb{P}$, and the associated integral, or expectation, will be denoted by $\mathbb{E}$.

If $\mu \in P(\mathcal{X})$ and $\nu \in P(\mathcal{Y})$ are given, then $\Pi(\mu, \nu)$ is the set of all joint probability measures on $\mathcal{X} \times \mathcal{Y}$ whose marginals are μ and ν.

If $\pi(dx\, dy)$ is a probability measure in two variables x and y, the conditional law of x given y will be denoted by $\pi(dx|y)$; this is a measurable function of y with values in the set of probability measures in the variable x; it is obtained by disintegrating π along its y-marginal.

Notation specific to optimal transport

$C(\mu, \nu)$ is the optimal (total) cost between μ and ν, see p. 66. It implicitly depends on the choice of a cost function $c(x, y)$.

For any $p \in [1, +\infty)$, W_p is the Wasserstein distance of order p, see Definition 6.1; and $P_p(\mathcal{X})$ is the Wasserstein space of order p, i.e. the set of probability measures with finite moments of order p, equipped with the distance W_p, see Definition 6.3.

$P_c(\mathcal{X})$ is the set of probability measures on $\mathcal{X}$ with compact support.

If a reference measure ν on $\mathcal{X}$ is specified, then $P^{\mathrm{ac}}(\mathcal{X})$ (resp. $P_p^{\mathrm{ac}}(\mathcal{X})$, $P_c^{\mathrm{ac}}(\mathcal{X})$) stands for those elements of $P(\mathcal{X})$ (resp. $P_p(\mathcal{X})$, $P_c(\mathcal{X})$) which are absolutely continuous with respect to ν.

c-convex functions are introduced in Definition 5.2.

$\mathcal{DC}_N$ is the displacement convex class of order N (N plays the role of a dimension); this is a family of convex functions, defined on p. 273 and in Definition 17.1.

U_ν is a functional defined on $P(\mathcal{X})$; it depends on a convex function U and a reference measure ν on $\mathcal{X}$. This functional will be defined at various levels of generality, first in equation (15.2), then in Definition 29.1 and Theorem 30.4.

$U_{\pi,\nu}^{\beta}$ is another functional on $P(\mathcal{X})$, which involves not only a convex function U and a reference measure ν, but also a coupling π and a distortion coefficient β, which is a nonnegative function on $\mathcal{X} \times \mathcal{X}$: See again Definition 29.1 and Theorem 30.4.

$\beta_t^{(K,N)}$ is the notation for the distortion coefficients that will play a prominent role in these notes; they are defined in (14.60).

$\mathrm{CD}(K,N)$ means "curvature-dimension condition (K,N)", which morally means that the Ricci curvature is bounded below by K (a real number) and the dimension is bounded above by N (a real number which is not less than 1).

If $\pi(dx\,dy)$ is a coupling, then $\check{\pi}$ is the coupling obtained by swapping variables, that is $\pi(dy\,dx)$, or more rigorously, $S_{\#}\pi$, where $S(x,y) = (y,x)$.

Introduction

For a start, I shall recall in Chapter 1 some basic facts about couplings and changes of variables, including definitions, a short list of some famous couplings (Knothe-Rosenblatt coupling, Moser coupling, optimal coupling, etc.); and some important basic formulas about change of variables, conservation of mass, and linear diffusion equations.

In Chapter 2 I shall present, without detailed proofs, three applications of optimal coupling techniques, providing a flavor of the kind of applications that will be considered later.

Finally, Chapter 3 is a short historical perspective about the foundations and development of optimal coupling theory.

1

Couplings and changes of variables

Couplings are very well-known in all branches of probability theory, but since they will occur again and again in this course, it might be a good idea to start with some basic reminders and a few more technical issues.

Definition 1.1 (coupling). *Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two probability spaces. Coupling μ and ν means constructing two random variables X and Y on some probability space Ω, in such a way that* $\text{law}\,(X) = \mu$, $\text{law}\,(Y) = \nu$. *The couple (X, Y) is called a coupling of (μ, ν). By abuse of language, one often says that the law of (X, Y) is a coupling of (μ, ν).*

If μ and ν are the only laws in the problem, then without loss of generality one may choose $\Omega = \mathcal{X} \times \mathcal{Y}$. In a more measure-theoretical formulation, coupling μ and ν means constructing a measure π on $\mathcal{X} \times \mathcal{Y}$ such that π admits μ and ν as **marginals** on $\mathcal{X}$ and $\mathcal{Y}$ respectively. The following three statements are equivalent ways to rephrase that marginal condition:

- $(\text{proj}_{\mathcal{X}})_{\#}\pi = \mu$, $(\text{proj}_{\mathcal{Y}})_{\#}\pi = \nu$, where $\text{proj}_{\mathcal{X}}$ and $\text{proj}_{\mathcal{Y}}$ respectively stand for the projection maps $(x, y) \longmapsto x$ and $(x, y) \longmapsto y$;
- For all measurable sets $A \subset \mathcal{X}$, $B \subset \mathcal{Y}$, $\pi[A \times \mathcal{Y}] = \mu[A]$, $\pi[\mathcal{X} \times B] = \nu[B]$;
- For all integrable (resp. nonnegative) measurable functions φ, ψ on $\mathcal{X}, \mathcal{Y}$,

$$\int_{\mathcal{X}\times\mathcal{Y}} \big(\varphi(x) + \psi(y)\big)\, d\pi(x, y) = \int_{\mathcal{X}} \varphi\, d\mu + \int_{\mathcal{Y}} \psi\, d\nu.$$

A first remark about couplings is that they always exist: at least there is the **trivial coupling**, in which the variables X and Y are **independent** (so their joint law is the tensor product $\mu \otimes \nu$). This can hardly be called a coupling, since the value of X does not give any information about the value of Y. Another extreme is when all the information about the value of Y is contained in the value of X, in other words Y is just a function of X. This motivates the following definition (in which X and Y do not play symmetric roles).

Definition 1.2 (deterministic coupling). *With the notation of Definition 1.1, a coupling (X, Y) is said to be deterministic if there exists a measurable function $T : \mathcal{X} \to \mathcal{Y}$ such that*

$$Y = T(X).$$

To say that (X, Y) is a deterministic coupling of μ and ν is strictly equivalent to any one of the four statements below:

- (X, Y) is a coupling of μ and ν whose law π is concentrated on the *graph* of a measurable function $T : \mathcal{X} \to \mathcal{Y}$;
- X has law μ and $Y = T(X)$, where T *pushes* μ *forward* to ν: $T_\# \mu = \nu$.
- X has law μ and $Y = T(X)$, where T is a **change of variables** from μ to ν: for all integrable (resp. nonnegative) functions φ,

$$\int_{\mathcal{Y}} \varphi(y)\, d\nu(y) = \int_{\mathcal{X}} \varphi\big(T(x)\big)\, d\mu(x); \tag{1.1}$$

- $\pi = (\mathrm{Id}\,, T)_\# \mu$.

The map T appearing in all these statements is the same and is uniquely defined μ-almost surely (when the joint law of (X, Y) has been fixed). The converse is true: If T and $\widetilde{T}$ coincide μ-almost surely, then $T_\# \mu = \widetilde{T}_\# \mu$. It is common to call T the **transport map**: Informally, one can say that T transports the mass represented by the measure μ, to the mass represented by the measure ν.

Unlike couplings, deterministic couplings do not always exist: Just think of the case when μ is a Dirac mass and ν is not. But there may also be infinitely many deterministic couplings between two given probability measures.

Some famous couplings

Here below are some of the most famous couplings used in mathematics — of course the list is far from complete, since everybody has his or her own preferred coupling technique. Each of these couplings comes with its own natural setting; this variety of assumptions reflects the variety of constructions. (This is a good reason to state each of them with some generality.)

1. The **measurable isomorphism:** Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be Polish (i.e. complete, separable, metric) probability spaces without atom (i.e. no single point carries a positive mass). Then there exists a (nonunique) measurable bijection $T : \mathcal{X} \to \mathcal{Y}$ such that $T_\# \mu = \nu$, $(T^{-1})_\# \nu = \mu$. In that sense, all atomless Polish probability spaces are isomorphic, and, say, isomorphic to the space $\mathcal{Y} = [0, 1]$ equipped with the Lebesgue measure. Powerful as that theorem may seem, in practice the map T is very singular; as a good exercise, the reader might try to construct it "explicitly", in terms of cumulative distribution functions, when $\mathcal{X} = \mathbb{R}$ and $\mathcal{Y} = [0, 1]$ (issues do arise when the density of μ vanishes at some places). Experience shows that it is quite easy to fall into logical traps when working with the measurable isomorphism, and my advice is to never use it.

2. The **Moser mapping:** Let $\mathcal{X}$ be a smooth compact Riemannian manifold with volume $\mathrm{vol}\,$, and let f, g be Lipschitz continuous positive probability densities on $\mathcal{X}$; then there exists a deterministic coupling T of $\mu = f\,\mathrm{vol}\,$ and $\nu = g\,\mathrm{vol}\,$. On the plus side, there is a somewhat explicit representation of that mapping, and it is as smooth as can be: if f, g are $C^{k,\alpha}$ then T is $C^{k+1,\alpha}$. The formula is given in an Appendix at the end of this chapter. The same construction works in $\mathbb{R}^n$ provided that f and g decay fast enough at infinity; and it is robust enough to accomodate for variants.

3. The **increasing rearrangement** on $\mathbb{R}$. Let μ, ν be two probability measures on $\mathbb{R}$, define their cumulative distribution functions by

$$F(x) = \int_{-\infty}^{x} d\mu, \qquad G(y) = \int_{-\infty}^{y} d\nu.$$

Further define their right-continuous inverses by

$$F^{-1}(t) := \inf\Big\{x \in \mathbb{R};\ F(x) > t\Big\}; \qquad G^{-1}(t) := \inf\Big\{y \in \mathbb{R};\ G(y) > t\Big\};$$

and set

$$T = G^{-1} \circ F.$$

If μ does not have atoms, then $T_{\#}\mu = \nu$. This rearrangement is quite simple, explicit, as smooth as can be, and has good geometric properties.

4. The **Knothe-Rosenblatt rearrangement** in $\mathbb{R}^n$. Let μ and ν be two probability measures on $\mathbb{R}^n$, such that μ is absolutely continuous with respect to Lebesgue measure. Then define a coupling of μ and ν as follows.

   Step 1: Take the marginal on the first variable: this gives probability measures $\mu_1(dx_1)$, $\nu_1(dy_1)$ on $\mathbb{R}$, with μ_1 atomless. Define $y_1 = T_1(x_1)$ by the formula for the increasing rearrangement of μ_1 into ν_1.

   Step 2: Now take the marginal on the first two variables and disintegrate it with respect to the first variable: This gives probability measures $\mu_2(dx_1\, dx_2) = \mu_1(dx_1)\, \mu_2(dx_2|x_1)$, $\nu_2(dy_1\, dy_2) = \nu_1(dy_1)\, \nu_2(dy_2|y_1)$. Then, for each given $y_1 \in \mathbb{R}$, set $y_1 = T_1(x_1)$, and define $y_2 = T_2(x_2; x_1)$ by the formula for the increasing rearrangement of $\mu_2(dx_2|x_1)$ into $\nu_2(dy_2|y_1)$.

   Then repeat the construction, adding variables one after the other and defining $y_3 = T_3(x_3; x_1, x_2)$; etc. After n steps, this produces a map $y = T(x)$ which transports μ to ν, and in practical situations might be computed explicitly with little effort. Moreover, the Jacobian matrix of the change of variables T is (by construction) upper triangular with positive entries of the diagonal; this makes it suitable for various geometric applications. On the negative side, this mapping does not satisfy many interesting intrinsic properties; it is not invariant under isometries of $\mathbb{R}^n$, not even under relabelling of coordinates.

5. The **Holley coupling** on a lattice. Let μ and ν be two discrete probabilities on a finite lattice Λ, say $\{0,1\}^N$, equipped with the natural partial ordering ($x \le y$ if $x_n \le y_n$ for all n). Assume that

$$\forall x, y \in \Lambda, \qquad \mu[\inf(x,y)]\ \nu[\sup(x,y)] \ \ge\ \mu[x]\, \nu[y]. \tag{1.2}$$

   Then there exists a coupling (X, Y) of (μ, ν) with $X \le Y$. The situation above appears in a number of problems in statistical mechanics, in connection with the so-called FKG inequalities. Inequality (1.2) intuitively says that ν puts more mass on large values than μ.

6. **Probabilistic representation formulas** for solutions of partial differential equations. There are hundreds of them (if not thousands), representing solutions of diffusion, transport or jump processes as the laws of various deterministic or stochastic processes. Some of them are recalled later in this chapter.

7. The **exact coupling** of two stochastic processes, or Markov chains. Two realizations of a stochastic process are started at initial time, and when they happen to be in the same state at some time, they are merged: From that time on, they follow the same path and accordingly, have the same law. For two Markov chains which are started

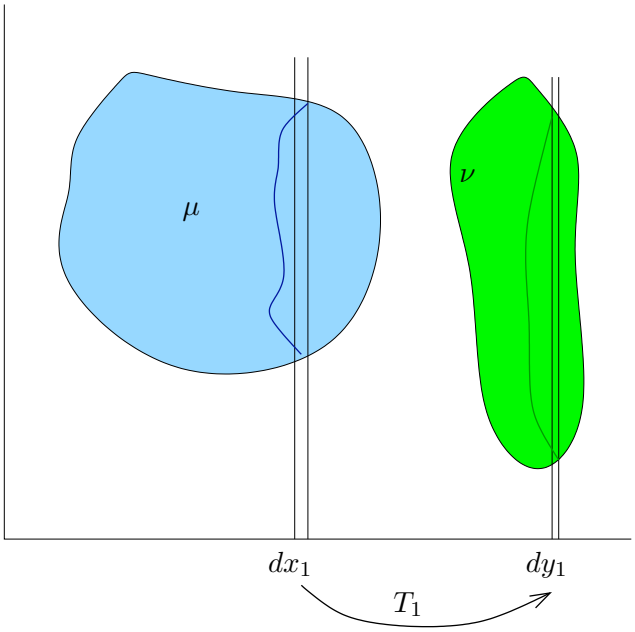


Fig. 1.1. Second step in the construction of the Knothe-Rosenblatt map: After the correspondance $x_1 \to y_1$ has been determined, the conditional probability of x_2 (seen as a one-dimensional probability on a small "slice" of width dx_1) can be transported to the conditional probability of y_2 (seen as a one-dimensional probability on a slice of width dy_1).

independently, this is called the **classical coupling**. There are many variants with important differences which are all intended to make two trajectories close to each other after some time: the **Ornstein coupling**, the **ε-coupling** (in which one requires the two variables to be close, rather than to occupy the same state), the **shift-coupling** (in which one allows an additional time-shift), etc.

8. The **optimal coupling** or **optimal transport**. Here one introduces a **cost function** $c(x, y)$ on $\mathcal{X} \times \mathcal{Y}$, that can be interpreted as the infinitesimal work needed to move one unit of mass from location x to location y. Then one considers the **Monge-Kantorovich minimization problem**

$$\inf \ \mathbb{E}\, c(X, Y),$$

where the pair (X, Y) runs over all possible couplings of (μ, ν); or equivalently, in terms of measures,

$$\inf \int_{\mathcal{X} \times \mathcal{Y}} c(x, y)\, d\pi(x, y),$$

where the infimum runs over all joint probability measures π on $\mathcal{X} \times \mathcal{Y}$ with marginals μ and ν. Such joint measures are called **transport plans** (or transportation plans), and optimal ones are called **optimal transport plans**.

Of course, the solution of the Monge-Kantorovich problem depends on the cost function c. The cost function and the probability spaces here can be very general, and some nontrivial results about the Monge-Kantorovich problem can be obtained as soon, say, c is lower semi-continuous and $\mathcal{X}, \mathcal{Y}$ are Polish spaces. Even the apparently trivial choice $c(x, y) = 1_{x \neq y}$ appears in the probabilistic interpretation of total variation:

$$\|\mu - \nu\|_{TV} = 2 \inf \Big\{ \mathbb{E}\, 1_{X \neq Y}; \quad \text{law}\,(X) = \mu, \ \text{law}\,(Y) = \nu \Big\}.$$

Cost functions valued in $\{0, 1\}$ also appear naturally in Strassen's duality theorem.

Under certain assumptions one can guarantee that the optimal coupling really is *deterministic*; the search of deterministic optimal couplings is called the **Monge problem**. A solution of the Monge problem yields a plan to transport the mass at minimal cost with a recipe that associates to each point x a single point y. ("*No mass shall be split*".) To force the existence of solutions to the Monge problem, two kinds of assumptions are natural: First, c should "vary enough" in some sense (think that the constant cost function will allow for arbitrary minimizers), and secondly, μ should enjoy some regularity property (at least Dirac masses should be ruled out!)

Here is a typical result: If $c(x, y) = |x - y|^2$ in the Euclidean space μ is absolutely continuous with respect to Lebesgue measure, and μ, ν have finite moments of order 2, then there is a unique optimal Monge coupling between μ and ν. More general statements will be established in Chapter 10.

Optimal couplings enjoy many nice properties:

(i) They naturally arise in many problems coming from economics, physics, partial differential equations or geometry (by the way, the increasing rearrangement and the Holley coupling can be seen as particular cases of optimal transport);

(ii) They are quite stable with respect to perturbations;

(iii) They encode good geometric information, if the cost function c is defined in terms of the underlying geometry;

(iv) They exist in smooth as well as nonsmooth settings;

(v) They come with a rich structure: an **optimal cost** functional (the value of the minimum defining the Monge-Kantorovich problem); a **dual variational problem**; and, under adequate structure conditions, a continuous **interpolation**.

On the negative side, it is important to be warned that optimal transport is in general not so smooth. There are known counterexamples which put limits on the regularity that one can expect from it, even for very nice cost functions.

All these issues will be discussed again and again in the sequel.

Gluing

If Z is a function of Y and Y is a function of X, then of course Z is a function of X. Something of this still remains true in the setting of nondeterministic couplings, under quite general assumptions.

Gluing Lemma. *Let $(\mathcal{X}_i, \mu_i)$, $i = 1, 2, 3$, be Polish probability spaces. If (X_1, X_2) is a coupling of (μ_1, μ_2) and (Y_2, Y_3) is a coupling of (μ_2, μ_3), then one can construct a triple of random variables (Z_1, Z_2, Z_3) such that (Z_1, Z_2) has the same law as (X_1, X_2) and (Z_2, Z_3) has the same law as (Y_2, Y_3).*

It is simple to understand why this is called "gluing lemma": if π_{12} stands for the law of (X_1, X_2) on $\mathcal{X}_1 \times \mathcal{X}_2$ and π_{23} stands for the law of (X_2, X_3) on $\mathcal{X}_2 \times \mathcal{X}_3$, then to construct the joint law π_{123} of (Z_1, Z_2, Z_3) one just has to *glue* π_{12} and π_{23} along their common marginal μ_2. In a slightly informal writing: Disintegrate π_{12} and π_{23} as

$$\pi_{12}(dx_1\, dx_2) = \pi_{12}(dx_1|x_2)\, \mu_2(dx_2), \qquad \pi_{23}(dx_2\, dx_3) = \pi_{23}(dx_3|x_2)\, \mu_2(dx_2),$$

and then reconstruct π_{123} as

$$\pi_{123}(dx_1\, dx_2\, dx_3) = \pi_{12}(dx_1|x_2)\, \mu_2(dx_2)\, \pi_{23}(dx_3|x_2).$$

Change of Variables Formula

When one writes the formula for change of variables, say in $\mathbb{R}^n$ or on a Riemannian manifold, a Jacobian term appears, and one has to be careful about two things: the change of variables should be *injective* (otherwise, reduce to a subset where it is injective, or take the multiplicity into account); and it should be somewhat smooth. It is classical to write these formulas when the change of variables is continuously differentiable, or at least Lipschitz:

Change of Variables Formula. *Let M be an n-dimensional Riemannian manifold with a C^1 metric, let μ_0, μ_1 be two probability measures on M, and let $T : M \to M$ be a measurable function such that $T_\# \mu_0 = \mu_1$. Let ν be a reference measure, of the form $\nu(dx) = e^{-V(x)} \, \mathrm{vol}\,(dx)$, where V is continuous and* vol *is the volume (or n-dimensional Hausdorff) measure. Further assume that*

(i) $\mu_0(dx) = \rho_0(x)\, \nu(dx)$ and $\mu_1(dy) = \rho_1(y)\, \nu(dy)$;

(ii) T is injective;

(iii) T is locally Lipschitz.

Then, μ_0-almost surely,

$$\rho_0(x) = \rho_1(T(x))\, \mathcal{J}_T(x), \tag{1.3}$$

where $\mathcal{J}_T(x)$ is the Jacobian determinant of T at x, defined by

$$\mathcal{J}_T(x) := \lim_{\varepsilon \downarrow 0} \frac{\nu[T(B_\varepsilon(x))]}{\nu[B_\varepsilon(x)]}. \tag{1.4}$$

The same holds true if T is only defined on the complement of a μ_0-negligible set, and satisfies properties (ii) and (iii) on its domain of definition.

Remark 1.3. When ν is just the volume measure, $\mathcal{J}_T(x)$ coincides with the usual Jacobian determinant, which in the case $M = \mathbb{R}^n$ is the absolute value of the determinant of the Jacobian matrix ∇T. Since V is continuous, it is almost immediate to deduce the statement with an arbitrary V from the statement with $V = 0$ (this amounts to multiply $\rho_0(x)$ by $e^{V(x)}$, $\rho_1(y)$ by $e^{V(y)}$, $\mathcal{J}_T(x)$ by $e^{V(x) - V(T(x))}$).

Remark 1.4. There is a more general framework, namely **approximate differentiability**. A function T on an n-dimensional Riemannian manifold is said to be approximately differentiable at x if there exists a function $\widetilde{T}$, differentiable at x, such that the set $\{\widetilde{T} \neq T\}$ has zero density at x, meaning

$$\lim_{r \to 0} \frac{\mathrm{vol}\left[\left\{x \in B_r(x);\ \ T(x) \neq \widetilde{T}(x)\right\}\right]}{\mathrm{vol}\,[B_r(x)]} = 0.$$

It turns out that, roughly speaking, an approximately differentiable map can be replaced, up to neglecting a small set, by a Lipschitz map (this is a kind of differentiable version of Lusin's theorem). So one can prove the Jacobian formula for an approximately differentiable map by approximating it with a sequence of Lipschitz maps.

Approximate differentiability is obviously a local property; it holds true if the distributional derivative of T is a locally integrable function, or even a locally finite measure. Even if I shall not need them in this course, such singular changes of variables actually arise rather naturally in the study of optimal transportation. So it might be useful to know that the Theorem of Change of Variables Formula still holds true if (iii) is replaced by

(iii') T is approximately differentiable.

Conservation of Mass Formula

The single most important theorem of change of variables arising in continuum physics might be the one resulting from the **conservation of mass** formula,

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \xi) = 0. \tag{1.5}$$

Here $\rho = \rho(t,x)$ stands for the density of a system of particles at time t and position x; $\xi = \xi(t,x)$ for the velocity field at time t and position x; and $\nabla\cdot$ stands for the divergence operator (the trace of the Jacobian matrix with respect to the x variable). Once again, the natural setting for this equation is a Riemannian manifold M.

It will be useful to work with particle densities $\mu_t(dx)$ (that are not necessarily absolutely continuous) and rewrite (1.5) as

$$\frac{\partial \mu}{\partial t} + \nabla \cdot (\mu \xi) = 0,$$

where the divergence operator is defined by duality against continuously differentiable functions with compact support:

$$\int_M \varphi \, \nabla \cdot (\mu \xi) = - \int_M (\xi \cdot \nabla \varphi) \, d\mu.$$

The formula of conservation of mass is an **Eulerian** description of the physical world, which means that the unknowns are fields. The next theorem links it with the **Lagrangian** description, in which everything is expressed in terms of particle trajectories, that are integral curves of the velocity field:

$$\xi\big(t, T_t(x)\big) = \frac{d}{dt} T_t(x). \tag{1.6}$$

If ξ is (locally) Lipschitz continuous, then the Cauchy-Lipschitz theorem guarantees the existence of a flow T_t locally defined on a maximal time interval, and itself locally Lipschitz in both arguments t and x. Then, for each t the map T_t is a local diffeomorphism onto its image. But the formula of conservation of mass also holds true without any regularity assumption on ξ; one should only keep in mind that if ξ is not Lipschitz, then a solution of (1.6) is not uniquely determined by its value at time 0, so $x \longmapsto T_t(x)$ is not necessarily uniquely defined. On the other hand, it makes sense to consider *random* solutions of (1.6).

Mass Conservation Formula. *Let M be a C^1 manifold, $T \in (0, +\infty]$ and let $\xi(t,x)$ be a (measurable) velocity field on $[0,T) \times M$. Let $(\mu_t)_{0 \le t < T}$ be a time-dependent family of probability measures on M (continuous in time for the weak topology), such that*

$$\int_0^T |\xi(t,x)| \, \mu_t(dx) \, dt < +\infty.$$

Then, the following two statements are equivalent:

(i) $\mu = \mu_t(dx)$ is a weak solution of the linear (transport) partial differential equation

$$\partial_t \mu + \nabla_x \cdot (\mu \xi) = 0$$

on $[0,T) \times M$;

(ii) μ_t is the law at time t of a random solution $T_t(x)$ of (1.6).

If moreover ξ is locally Lipschitz, then $(T_t)_{0 \le t < T}$ defines a deterministic flow, and statement (ii) can be rewritten

(ii') $\mu_t = (T_t)_\# \mu_0$.

Diffusion Formula

The final reminder in this chapter is very well-known and related to Itô's formula; it was discovered independently (in the Euclidean context) by Bachelier, Einstein and Smoluchowski at the beginning of the twentieth century. It requires a bit more regularity than the Conservation of Mass Formula. The natural assumptions on the phase space are in terms of *Ricci curvature*, a concept which will play an important role in these notes. For the reader who has no idea what Ricci curvature means, it is sufficient to know that the theorem below applies when M is either $\mathbb{R}^n$, or a compact manifold with a C^2 metric. By convention, B_t denotes the "standard" Brownian motion on M with identity covariance matrix.

Diffusion Theorem. *Let M be a Riemannian manifold with a C^2 metric, such that the Ricci curvature tensor of M is uniformly bounded below, and let $\sigma(t,x) : T_xM \to T_xM$ be a twice differentiable linear mapping on each tangent space. Let X_t stand for the solution of the stochastic differential equation*

$$dX_t = \sqrt{2}\,\sigma(t, X_t)\, dB_t \qquad (0 \le t < T). \tag{1.7}$$

Then the following two statements are equivalent:

(i) $\mu = \mu_t(dx)$ is a weak solution of the linear (diffusion) partial differential equation

$$\partial_t \mu = \nabla_x \cdot \Big((\sigma\sigma^*) \nabla_x \mu \Big)$$

on $M \times [0,T)$, where σ^* stands for the transpose of σ;

(ii) $\mu_t = \text{law}\,(X_t)$ for all $t \in [0,T)$, where X_t solves (1.7).

Example 1.5. In $\mathbb{R}^n$, the solution of the heat equation with initial datum δ_0 is the law of $X_t = \sqrt{2}\, B_t$ (Brownian motion sped up by a factor $\sqrt{2}$).

Remark 1.6. Actually, there is a finer criterion for the diffusion equation to hold true: it is sufficient that the Ricci curvature at point x be bounded below by $-Cd(x_0,x)^2 g_x$, where g_x is the metric at point x and x_0 is an arbitrary reference point. The exponent 2 here is sharp.

Exercise 1.7. Let M be a smooth compact manifold, equipped with its standard reference volume, and let ρ_0 be a smooth positive probability density on M. Let $(\rho_t)_{t \ge 0}$ be the solution of the heat equation

$$\partial_t \rho = \Delta \rho.$$

Use (ρ_t) to construct a *deterministic* coupling of ρ_0 and ρ_1.

Hint: Rewrite the heat equation in the form of an equation of conservation of mass.

Appendix: Moser's coupling

In this Appendix I shall advertise for Moser's technique for coupling smooth positive probability measures; it is simple, elegant and powerful, and plays a prominent role in geometry. It is not limited to compact manifolds, but does require assumptions about the behavior at infinity.

Here I shall explain the method in the case when M is a smooth n-dimensional Riemannian manifold, equipped with a reference probability measure $\nu(dx) = e^{-V(x)}\,\text{vol}\,(dx)$,

and V is at least continuously differentiable. So let $\mu_0 = \rho_0\,\nu$, $\mu_1 = \rho_1\,\nu$ be two probability measures on M, and assume for simplicity that ρ_0, ρ_1 are bounded below by a constant $K > 0$. Further assume that ρ_0 and ρ_1 are locally Lipschitz, and that the equation

$$(\Delta - \nabla V \cdot \nabla) u = \rho_0 - \rho_1$$

can be solved for some $u \in C^{1,1}_{\mathrm{loc}}(M)$ (that is, ∇u is locally Lipschitz). Then, define a locally Lipschitz vector field

$$\xi(t,x) = \frac{\nabla u(x)}{(1-t)\rho_0(x) + t\rho_1(x)},$$

with associated flow $(T_t(x))_{0 \le t \le 1}$, and a family $(\mu_t)_{0<t<1}$ of probability measures by

$$\mu_t = (1-t)\mu_0 + t\mu_1.$$

It is easy to check that

$$\partial_t \mu = (\rho_1 - \rho_0)\,\nu,$$

$$\nabla \cdot (\mu_t \xi(t,\cdot)) = \nabla \cdot (\nabla u\, e^{-V}\,\mathrm{vol}\,) = e^{-V}(\Delta u - \nabla V \cdot \nabla u)\,\mathrm{vol}\, = (\rho_0 - \rho_1)\,\nu.$$

So μ_t satisfies the formula of conservation of mass, therefore $\mu_t = (T_t)_{\#}\mu_0$. In particular, T_1 provides a deterministic coupling of μ_0 and μ_1.

In the case when M is compact and $V = 0$, the above construction works if ρ_0 and ρ_1 are Lipschitz continuous and positive. Indeed, the solution of $\Delta u = \rho_0 - \rho_1$ will be of class $C^{2,\alpha}$ for all $\alpha \in (0,1)$, and in particular ∇u will be of class C^1. In more general situations, things might depend on the regularity of V, and its behavior at infinity.

Bibliographical Notes

An excellent general reference book for the "classical theory" of couplings is the monograph by Thorisson [352]. There one can find an exhaustive treatment of classical couplings of Markov chains or stochastic processes, such as ε-coupling, shift-coupling, Ornstein coupling. As for the classical theory of optimal couplings, a rather exhaustive account can be found in the two volumes by Rachev and Rüschendorf [306]. This includes in particular the theory of optimal coupling on the real line with a convex cost function, which can be treated in a simple and direct manner [306, Section 3.1]. (In [365], for the sake of consistency of the presentation I treated optimal coupling on $\mathbb{R}$ as a particular case of optimal coupling on $\mathbb{R}^n$, however this has the drawback to involve subtle arguments.)

The Knothe-Rosenblatt coupling was introduced in 1952 by Rosenblatt [314], who suggested that it might be useful to "normalize" statistical data before applying a statistical test. In 1957, Knothe [228] rediscovered it for applications to the theory of convex bodies. It is quite likely that other people have discovered this coupling independently.

It was in 1965 that Moser proved his coupling theorem, for smooth compact manifolds without boundaries [279]. Noncompact manifolds were later considered by Greene and Shiohama [190]. Moser himself also worked with Dacorogna on the more delicate case where the domain is an open set with boundary, and the transport is required to fix the boundary [126].

Strassen's duality theorem is discussed e.g. in [365, Section 1.4].

The Gluing Lemma is due to several authors, starting with Vorob'ev in 1962 for finite sets. The modern formulation seems to have emerged around 1980, independently by Berkes and Philipp [46], Thorisson, Kallenberg, and maybe others. Refinements were discussed e.g. by de Acosta [128, Theorem A.1] (for marginals indexed by an arbitrary

set) or Thorisson [352, Theorem 5.1]; see also the bibliographic comments in [146, p. 20]. For a proof of the statement in these notes, it is sufficient to consult Dudley [146, Theorem 1.1.10], or [365, Lemma 7.6]. A comment about terminology: I like the word "gluing" which seems to give a good intuition of the proof, but many authors just talk about "composition" of plans.

The formula of changes of variables can be found in many textbooks for C^1 or Lipschitz change of variables, see e.g. Evans and Gariepy [156, Chapter 3]. The generalization to approximately differentiable maps is explained in Ambrosio, Gigli and Savaré [15, Section 5.5]. Such a generality is interesting in the context of optimal transportation, where changes of variables are often very rough (say BV, which means of bounded variation). In that context however, there is more structure: For instance, changes of variables will typically be given by the gradient of a convex function in $\mathbb{R}^n$, and on such a map one knows slightly more than on a general BV function, because convex functions are twice differentiable almost everywhere in the sense of Alexandrov. McCann [267] used this property to prove, by slightly more elementary means, the change of variables formula for a gradient of convex function. The proof is reproduced in [365, Theorem 4.8]. It was later generalized by Cordero-Erausquin, McCann and Schmuckenschläger to Riemannian manifolds [118], a case which again can be treated either as part of the general result about BV changes of variables, or with the help of almost everywhere second derivatives of semi-concave functions.

The formula of conservation of mass is also called the method of characteristics for linear transport equations, and is described in a number of textbooks in partial differential equations, at least when the driving vector field is Lipschitz, see for instance Evans [153, Section 3.2]. An essentially equivalent statement is proven in [365, Theorem 5.34]. Treating vector fields that are only assumed to be locally Lipschitz is not so easy, see Ambrosio, Gigli and Savaré [15, Section 8.1].

The Lipschitz condition can be relaxed into a $W^{1,p}$ or even a BV condition, but then the flow is determined only almost everywhere, and this becomes an extremely subtle problem, which has been studied by many authors since the pioneering work of DiPerna and Lions at the beginning of the nineties. See Ambrosio [12] for recent progress and references. The version which is stated in these notes, with no regularity assumption, is due to Ambrosio and carefully proved in [15, Section 8.1]. In spite of its appealing and relatively natural character (especially in a probabilistic perspective), this is a very recent research result. Note that, if $T_t(x)$ is not uniquely determined by x, then the solution to the conservation equation starting with a given probability measure might admit several solutions.

A recent work by Lisini [244] addresses a generalization of the formula of conservation of mass in the setting of general Polish spaces. Of course, without any regularity assumption on the space it is impossible to speak of vector fields and partial differential equations; but it is still possible to consider paths in the space of probability measures, and random curves. Lisini's results are most naturally expressed in the language of optimal transport distances; see the bibliographical notes for Chapter 7.

The Diffusion Formula can be obtained as a simple consequence of the Itô formula, which in the Euclidean setting can be found in any textbook on stochastic differential equations, e.g. [285]. (Let me note that this was recently the hundredth anniversary of its discovery by Einstein [150].) Fascinating tales about the Brownian motion can be read in Nelson's unconventional book [282], especially Chapters 1–4. For the much more subtle Riemannian setting, one may consult Stroock [334], Hsu [216] and the references therein.

The Brownian motion on a smooth Riemannian manifold is always well-defined, even if the manifold has a wild behavior at infinity (the construction of the Brownian motion is

purely local); but in the absence of a good control on the Ricci curvature, there might be several heat kernels, and the heat equation might not be uniquely solvable from a given initial datum. This corresponds to the possibility of a blow-up of the Brownian motion (i.e. the Brownian motion escapes to infinity) in finite time. All this was explained to me by Thalmaier. The sharp criterion $\mathrm{Ric} \geq -K[1 + d(x_0, x)^2]$ for avoiding blow-up of the heat equation is based on comparison theorems for Laplace operators. In the version stated here it is due to Ichihara [218]; see also the book by Hackenbroch and Thalmaier [203, p. 544]. Non-explosion criteria based on curvature have been studied also by Gaffney, Yau, Hsu, Karp and Li, Davies, Takeda, Sturm and Grigor'yan; for a detailed exposition, and many explanations, the reader can consult the survey by Grigor'yan [191, Section 9].

2

Three examples of coupling techniques

In this chapter I shall present three applications of coupling methods. The first one is classical and quite simple, the other two are more original but very typical of the topics that will be considered later in these notes. The proofs are extremely variable in difficulty and will only be sketched here; see the references in the bibliographical notes for details.

Convergence of the Langevin process

Consider a particle subject to the force induced by a potential $V \in C^1(\mathbb{R}^n; \mathbb{R})$, a friction and a random white noise agitation. If X_t stands for the position of the particle at time t, m for its mass, λ for the friction coefficient, k for the Boltzmann constant and T for the temperature of the heat bath, then the Newton equations can be written as

$$m \frac{d^2 X_t}{dt^2} = -\nabla V(X_t) - \lambda m \frac{dX_t}{dt} + \sqrt{kT} \frac{dB_t}{dt}, \tag{2.1}$$

where $(B_t)_{t \geq 0}$ is a standard Brownian motion. This is a second-order (stochastic) differential equation, so it should come with initial conditions for both the position X and the velocity $\dot{X}$.

Now consider a large cloud particles evolving independently, according to (2.1); the question is whether the distribution of particles will converge to a definite limit as $t \to \infty$. In other words: Consider the stochastic differential equation (2.1) starting from some initial distribution $\mu_0(dx\, dv) = \text{law}\,(X_0, \dot{X}_0)$, is it true that $\text{law}\,(X_t)$, or $\text{law}\,(X_t, \dot{X}_t)$, will converge to some given limit law as $t \to \infty$?

Obviously, to solve this problem one has to make some assumptions on the potential V, which should prevent the particles from all escaping at infinity; for instance, we can make the very strong assumption that V is uniformly convex, i.e. there exists $K > 0$ such that the Hessian $\nabla^2 V$ satisfies $\nabla^2 V \geq K I_n$. Some assumptions on the initial distribution might also be needed; for instance, it is natural to assume that the Hamiltonian has finite expectation at initial time:

$$\mathbb{E}\left(V(X_0) + \frac{|\dot{X}_0|^2}{2} \right) < +\infty$$

Under these assumptions, it is true that there is exponential convergence to equilibrium, at least if V does not grow too wildly at infinity (for instance if the Hessian of V is also bounded above). However, I do not know of any simple method to prove this.

However, consider the limit where the friction coefficient is quite strong, and the motion of the particle is so slow that the acceleration term may be neglected in front of the others: Then, up to resetting units, equation (2.1) becomes

$$\frac{dX_t}{dt} = -\nabla V(X_t) + \frac{dB_t}{dt}, \tag{2.2}$$

which is often called a Langevin process. Now, to study the convergence of equilibrium for (2.2) there is an extremely simple solution by coupling. Consider another random position $(Y_t)_{t\geq 0}$ obeying the same equation as (2.2):

$$\frac{dY_t}{dt} = -\nabla V(Y_t) + \frac{dB_t}{dt}, \tag{2.3}$$

where the random realization of the Brownian motion is *the same* as in (2.2) (this is the coupling). The initial positions X_0 and Y_0 may be coupled in an arbitrary way, but it is possible to assume that they are independent. In any case, since they are driven by the same Brownian motion, X_t and Y_t will be correlated for $t > 0$.

Since B_t is not differentiable as a function of time, neither X_t nor Y_t is differentiable (equations (2.2) and (2.3) hold only in the sense of solutions of stochastic differential equations); but it is easily checked that $\alpha_t := X_t - Y_t$ is a continuously differentiable function of time, and

$$\frac{d\alpha_t}{dt} = -\big[\nabla V(X_t) - \nabla V(Y_t)\big],$$

so in particular

$$\frac{d}{dt}\,\frac{|\alpha_t|^2}{2} = -\Big\langle \nabla V(X_t) - \nabla V(Y_t),\ X_t - Y_t \Big\rangle \leq -K\big|X_t - Y_t\big|^2 = -K\,|\alpha_t|^2.$$

It follows by Gronwall's lemma that

$$|\alpha_t|^2 \leq e^{-2Kt}\,|\alpha_0|^2.$$

Assume for simplicity that $\mathbb{E}\,|X_0|^2$ and $\mathbb{E}\,|Y_0|^2$ are finite. Then

$$\mathbb{E}\,|X_t - Y_t|^2 \leq e^{-2Kt}\,\mathbb{E}\,|X_0 - Y_0|^2 \leq 2\,(\mathbb{E}\,|X_0|^2 + \mathbb{E}\,|Y_0|^2)\,e^{-2Kt}. \tag{2.4}$$

In particular, $X_t - Y_t$ converges to 0 almost surely, and this is independent of the distribution of Y_0.

This in itself would be essentially sufficient to guarantee the existence of a stationary distribution; but in any case, it is easy to check, by applying the Diffusion Formula, that

$$\nu(dy) = \frac{e^{-V(y)}}{Z}\,dy$$

(where $Z = \int e^{-V}$ is a normalization constant) is stationary: If $\mathrm{law}\,(Y_0) = \nu$, then also $\mathrm{law}\,(Y_t) = \nu$. Then (2.4) easily implies that $\mu_t := \mathrm{law}\,(X_t)$ converges weakly to ν; in addition, the convergence is exponentially fast.

Euclidean isoperimetry

Among all subsets of $\mathbb{R}^n$ with given surface, which one has the largest volume? To simplify the problem, let us assume that we are looking for a bounded open set $\Omega \subset \mathbb{R}^n$ with,

say, Lipschitz boundary $\partial\Omega$, that the measure of $|\partial\Omega|$ is given; then the problem is to maximize the measure of $|\Omega|$ itself For $\partial\Omega$ one should use the $(n-1)$-dimensional Hausdorff measure, and for Ω the n-dimensional Hausdorff measure, which of course is the same as the Lebesgue measure in $\mathbb{R}^n$.

It has been known, at least since ancient times, that the solution to this "isoperimetric problem" is the ball. A simple scaling argument shows that this statement is equivalent to the Euclidean **isoperimetric inequality**:

$$\frac{|\partial\Omega|}{|\Omega|^{\frac{n}{n-1}}} \geq \frac{|\partial B|}{|B|^{\frac{n}{n-1}}},$$

where B is any ball.

There are many many proofs of the isoperimetric inequality, and many refinements as well. It is less known that there is a proof by coupling.

Here is a sketch of the argument, forgetting about regularity issues. Let B be a ball such that $|\partial B| = |\partial\Omega|$. Consider a random point X distributed uniformly in Ω, and a random point Y distributed uniformly in B. Introduce the Knothe-Rosenblatt coupling of X and Y: This is a deterministic coupling of the form $Y = T(X)$, such that, at each $x \in \Omega$, the Jacobian matrix $\nabla T(x)$ is triangular with nonnegative diagonal entries. Since the law of X (resp. Y) has uniform density $1/|\Omega|$ (resp. $1/|B|$), the Change of Variables Formula writes

$$\forall x \in \Omega \qquad \frac{1}{|\Omega|} = \big(\det \nabla T(x)\big)\ \frac{1}{|B|}. \tag{2.5}$$

Since ∇T is triangular, its Jacobian determinant can be written $\det \nabla T = \prod \lambda_i$, and its divergence $\nabla \cdot T = \sum \lambda_i$, where the nonnegative numbers $(\lambda_i)_{1\leq i\leq n}$ are the eigenvalues of ∇T. Then the arithmetic-geometric inequality $(\prod \lambda_i)^{1/n} \leq (\sum \lambda_i)/n$ becomes

$$\big(\det \nabla T(x)\big)^{1/n} \leq \frac{\nabla \cdot T(x)}{n}.$$

Combining this with (2.5) results in

$$\frac{1}{|\Omega|^{1/n}} \leq \frac{(\nabla \cdot T)(x)}{n|B|^{1/n}}.$$

Integrate this over Ω and then apply the divergence theorem:

$$|\Omega|^{1-\frac{1}{n}} \leq \frac{1}{n|B|^{\frac{1}{n}}} \int_\Omega (\nabla \cdot T)(x)\, dx = \frac{1}{n|B|^{\frac{1}{n}}} \int_{\partial\Omega} (T \cdot \sigma)\, d\mathcal{H}^{n-1}, \tag{2.6}$$

where $\sigma : \partial\Omega \to \mathbb{R}^n$ is the unit outer normal to Ω and $\mathcal{H}^{n-1}$ is the $(n-1)$-dimensional Hausdorff measure (restricted to $\partial\Omega$). But T is valued in B, so $|T \cdot \sigma| \leq 1$, and (2.6) implies

$$|\Omega|^{1-\frac{1}{n}} \leq \frac{|\partial\Omega|}{n|B|^{\frac{1}{n}}}.$$

Since $|\partial\Omega| = |\partial B| = n|B|$, the right-hand side is actually $|B|^{1-\frac{1}{n}}$, so the volume of Ω is indeed bounded by the volume of B. This concludes the proof.

Open Problem 2.1. *Can one devise an optimal coupling between sets (in the sense of a coupling between the uniform probability measures on these sets) in such a way that the total cost of the coupling decreases under a some evolution converging to balls, such as mean curvature motion?*

Caffarelli's log concave perturbation theorem

The previous example was about transporting a set to another, now the present one is in some sense about transporting a whole space to another.

It is classical in geometry to compare a space $\mathcal{X}$ with a "model space" $\mathcal{M}$ that has nice properties and is, e.g., less curved than $\mathcal{X}$. A general idea is that certain inequalities which hold true on the model space can automatically be "transported" to $\mathcal{X}$. The theorem discussed in this section is a striking illustration of this idea.

Let F, G, H, J, L be nonnegative continuous functions on $\mathbb{R}$, with H and J nondecreasing, and let $\ell \in \mathbb{R}$. For a given measure μ on $\mathbb{R}^n$, let $\lambda[\mu]$ be the largest $\lambda \geq 0$ such that, for all Lipschitz functions $h : \mathbb{R}^n \to \mathbb{R}$,

$$\int_{\mathbb{R}^n} L(h)\, d\mu = \ell \quad \Longrightarrow \quad F\left(\int_{\mathbb{R}^n} G(h)\, d\mu\right) \leq \frac{1}{\lambda}\, H\left(\int_{\mathbb{R}^n} J(|\nabla h|)\, d\mu\right), \tag{2.7}$$

Functional inequalities of the form (2.7) are variants of Sobolev inequalities; many of them are well-known and useful. Caffarelli's theorem states that *they can only be improved by log-concave perturbation of the Gaussian distribution.* More precisely, if γ is the standard Gaussian measure and $\mu = e^{-v}\gamma$ is another probability measure, with v convex, then

$$\lambda[\mu] \geq \lambda[\gamma].$$

His proof is a simple consequence of the following remarkable fact, which I shall call **Caffarelli's log-concave perturbation theorem**: *If $d\mu/d\gamma$ is log-concave, then there exists a 1-Lipschitz change of variables from the measure γ to the measure μ.*

In other words, there is a deterministic coupling $\big(X,\ Y = \mathcal{C}(X)\big)$ of (γ, μ), such that $|\mathcal{C}(x) - \mathcal{C}(y)| \leq |x - y|$, or equivalently $|\nabla \mathcal{C}| \leq 1$. It follows in particular that

$$\big|\nabla(h \circ \mathcal{C})\big| \leq |(\nabla h) \circ \mathcal{C}|, \tag{2.8}$$

whatever the function h.

Now it is easy to understand why the existence of the map $\mathcal{C}$ implies the result about Sobolev inequalities: On one hand, the definition of change of variables implies

$$\int G(h)\, d\mu = \int G(h \circ \mathcal{C})\, d\gamma, \qquad \int L(h)\, d\mu = \int L(h \circ \mathcal{C})\, d\gamma;$$

on the other hand, by the definition of change of variables again, inequality (2.8) and the nondecreasing property of J,

$$\int J(|\nabla h|)\, d\mu = \int J\big(|\nabla h \circ \mathcal{C}|\big)\, d\gamma \geq \int J\big(|\nabla(h \circ \mathcal{C})|\big)\, d\gamma.$$

Thus, inequality (2.7) is indeed "transported" from the space $(\mathbb{R}^n, \gamma)$ to the space $(\mathbb{R}^n, \mu)$.

Bibliographical Notes

It is very classical to use coupling arguments to prove convergence to equilibrium for stochastic differential equations and Markov chains; many examples are described by Rachev and Rüschendorf [306] or Thorisson [352]. Actually, the standard argument found in textbooks to prove the convergence to equilibrium for a positive aperiodic ergodic Markov chain is a coupling argument (but the null case can also be treated in

that way, as I learnt from Thorisson). Optimal couplings are often well adapted to such situations, but definitely not the only ones to apply.

The coupling method is not limited to systems of independent particles, and sometimes works in presence of correlations, for instance if the law satisfies a nonlinear diffusion equation. This is exemplified in works by Tanaka [350] on the spatially homogeneous Boltzmann equation with Maxwell molecules (the core of Tanaka's argument is reproduced in my book [365, Section 7.5]), or some recent works [255, 102].

Recently, Cattiaux and Guillin [107] found a simple and elegant coupling argument to prove the exponential convergence for the law of the stochastic process

$$dX_t = \sqrt{2}\, dB_t - \widetilde{\mathbb{E}}\, \nabla V(X_t - \widetilde{X}_t)\, dt,$$

where $\widetilde{X}_t$ is an independent copy of X_t, the $\widetilde{\mathbb{E}}$ expectation only bears on $\widetilde{X}_t$, and V is assumed to be a uniformly convex C^1 symmetric potential on $\mathbb{R}^n$.

It is also classical to couple a system of particles with an auxiliary artificial system to study the limit when the number of particles becomes large. For the Vlasov equation in kinetic theory this was been done by Dobrushin [144] and Neunzert [284] several decades ago. (The proof is reproduced in Spohn [332, Chapter 5], and also suggested as an exercise in my book [365, Problem 14].) Later Sznitman used this strategy in a systematic way for the propagation of chaos, and made it very popular, see e.g. his work on the Boltzmann equation [343] or his Saint-Flour lecture notes [344] and the many references included.

In all these works, the "philosophy" is always the same: Introduce some nice coupling and see how it evolves in a certain asymptotic regime (say, either the time, or the number of particles, or both, go to infinity).

It is possible to treat the convergence to equilibrium for the complete system (2.1) by methods that are either analytic [140, 210, 366] or probabilistic [311, 261, 242], but all methods known to me are much delicate than the simple coupling argument which works for (2.2). It is certainly a nice open problem to find an elementary coupling argument that works for (2.1). (The arguments in the above-mentioned probabilistic proofs ultimately rely on coupling methods via theorems of convergence for Markov chains; but in a quite indirect way.)

Coupling techniques have also been used recently for proving rather spectacular uniqueness theorems for invariant measures in infinite dimension, see e.g. [149, 205, 204].

Classical references for the isoperimetric inequality and related topics are the books by Burago and Zalgaller [82], and Schneider [329]; and the survey by Osserman [286]. Knothe [228] had the idea to use a "coupling" method to prove geometric inequalities, and Gromov [275, Appendix] applied this method to prove the Euclidean isopetrimetric inequality. Trudinger [356] gave a closely related treatment of the same inequality and some of its generalizations, by means of a clever use of the Monge-Ampère equation (which more or less amounts to the construction of an optimal coupling with quadratic cost function). Cabré [86] found a surprising simplification of Trudinger's method, based on the solution of just a linear elliptic equation. The "proof" which I gave in this chapter is a variation on Gromov's argument (somewhat close to a recent work of Maggi and myself [253]); although it is not rigorous, there is no real difficulty in turning it into a full proof by using the same arguments as in the above-mentioned references. The reader may also consult Chapter 6 of my book [365], or the notes at the end of Chapters 18 and 21, for further links between optimal coupling and isoperimetric-type inequalities.

The construction of Caffarelli's map $\mathcal{C}$ is easy, at least conceptually: The optimal coupling of the Gaussian measure γ with the measure $\mu = e^{-v}\gamma$, when the cost function is the square of the Euclidean distance, will do the job. But proving that $\mathcal{C}$ is indeed 1-Lipschitz

is much more of a sport, and involves some techniques from nonlinear partial differential equations [91]. An idea of the core of the proof is explained in [365, Problem 13]. It would be nice to find a softer argument.

Üstünel pointed out to me that, if v is convex and symmetric ($v(-x) = v(x)$), then the Moser transport T from γ to $e^{-v}\gamma$ is contracting, in the sense that $|T(x)| \leq |x|$; it is not clear however that T would be 1-Lipschitz.

Caffarelli's theorem has many analytic and probabilistic applications, see e.g. [183, 207]. There is an infinite-dimensional version by Feyel and Üstünel [168], where the Gaussian measure is replaced by the Wiener measure.

3

The founding fathers of optimal transport

As many other research subjects in mathematics, the field of optimal transport was born several times. The first of these births occurred at the end of the eighteenth century, by ways of the French geometer Gaspard Monge.

Monge was born in 1746 under the French Ancient Régime. Because of his outstanding skills, military authorities tolerated him in a military training school from which he should have been excluded by his modest origin. He invented descriptive geometry all by his own, and the power of the method was so apparent that he was appointed professor at the age of 22, with the understanding that his theory would remain a military secret, for exclusive use of higher officers. He later served as a professor under several regimes (escaping a death sentence by the Terror, and becoming later one of Napoleon's closest friends), and taught at Ecole Normale Supérieure and École Polytechnique in Paris. Most of his work was devoted to geometry.

In 1781 he published one of his first famous works, *Mémoire sur la théorie des déblais et des remblais* (a "déblai" is an amount of material that is extracted from the earth or a mine; a "remblai" is a material that is input into a new construction). The problem considered by Monge is as follows: Assume you have a certain amount of soil, to extract from the ground and transport to places where it should be incorporated in a construction. The places where the material should be extracted, and the ones where it should be transported to, are all known. But the assignment has to be determined: To which destination should one send the material that has been extracted at a certain place? The answer does matter because transport is costly, and you want to minimize the total cost. Monge assumed that the transport cost of one unit of mass along a certain distance was given by the product of the mass by the distance.

Nowadays there is a Monge street in Paris, and therein one can find an excellent bakery called *Le Boulanger de Monge*. To acknowledge this, and to illustrate how Monge's problem can be recast in an economic perspective, I shall express the problem as follows. Consider a large number of bakeries, producing breads, that should be transported each morning to *cafés* where consumers will eat them. The amount of bread that can be produced at each bakery, and the amount that will be consumed at each café are known in advance, and can be modelled as probability measures (there is a "density of production" and a "density of consumption") on a certain space, which in our case would be Paris (equipped with the natural metric where the distance between two points is the length of the shortest path joining them). The problem is to find in practice where each unit of bread should go, in such a way as to minimize the total transport cost. So Monge's problem really is the search of an optimal coupling; and to be more precise, he was looking for a *deterministic* optimal coupling.

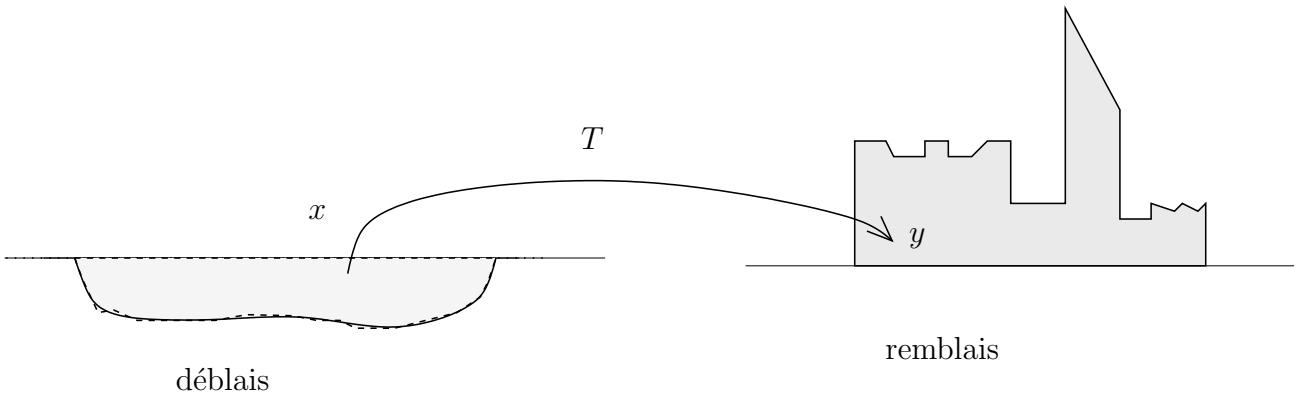


Fig. 3.1. Monge's problem of déblais and remblais

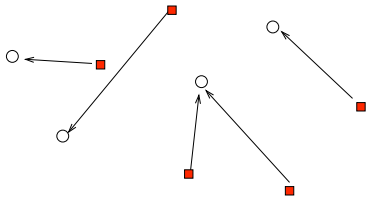


Fig. 3.2. Economic illustration of Monge's problem: squares stand for production units, circles for consumption places

Monge studied the problem in three dimensions for a continuous distribution of mass. Guided by his beautiful geometric intuition, he made the important observation that transport should go along straight lines that should be orthogonal to a family of surfaces. This study led him to the discovery of *lines of curvature*, a concept that by itself was a great contribution to the geometry of surfaces. His ideas were developed by Charles Dupin and later by Paul Appell. For nowadays' mathematical standards, all these arguments were flawed, yet it certainly would be worth looking up all these problems with modern tools.

Later Monge's problem was rediscovered by the Russian mathematician Leonid Kantorovich. Born in 1912, Kantorovich was a very gifted mathematician who made his reputation as a first-class researcher at the age of eighteen, and earned a professor position just as young as Monge. He worked in many areas of mathematics, with a strong taste for applications in economics, and later theoretical computer science. In 1938 a laboratory consulted him for the solution of a certain optimization problem, which he found out was representative of a whole class of linear problems arising in various areas of economics. Motivated by this discovery, he developed the tools of linear programming, that later became prominent in economics. He was awarded a joint Nobel prize in 1975 with Tjalling Koopmans "for their contributions to the theory of optimum allocation of resources".

In the case that is of direct interest for us, namely optimal coupling, Kantorovich stated and proved, by means of functional analytical tools, a duality theorem that would play a crucial role later. On the same occasion he devised a convenient notion of distance between probability measures: the distance between two measures should be the optimal transport cost from one to the other, if the cost is chosen as the distance function. This distance between probability measures is nowadays called the Kantorovich-Rubinstein distance, and has proven to be particularly flexible in many applications.

It is only several years after his main results that Kantorovich made the connection with Monge's work. The problem of optimal coupling has since then been called the **Monge-Kantorovich problem**.

All along the second half of the twentieth century, optimal coupling techniques and variants of the Kantorovich-Rubinstein distance (nowadays often called under the name of Wasserstein distances, or many other denominations) were used by statisticians and probabilists. The "basis" space could be finite-dimensional, or infinite-dimensional: For instance, optimal couplings give interesting notions of distance between probability measures on path spaces. A particularly interesting contribution is due to Tanaka, who in the seventies used such a distance to study the time-behavior of a simple variant of the Boltzmann equation.

During that time, *reparametrization* techniques (yet another word for change of variables) were used by many people working on inequalities involving volumes or integrals. Long later it would be understood that optimal transport often provides useful reparametrizations.

At the end of the eighties, three directions of research emerged independently and almost simultaneously, which completely reshaped the whole picture of optimal transport.

One of them was John Mather's work on Lagrangian dynamical systems. Action-minimizing curves are basic important objects in the theory of dynamical systems, and the construction of closed action-minimizing curves satisfying certain qualitative properties is a classical problem. By the end of the eighties, Mather found it convenient to study not only action-minimizing curves, but action-minimizing stationary *measures* in phase space. Mather's measures are a generalization of action-minimizing curves, and they solve a variational problem which in effect is a Monge-Kantorovich problem. Under some conditions on the Lagrangian, Mather proved a celebrated result according to which (roughly speaking) certain action-minimizing measures are automatically concentrated on Lipschitz graphs. As we shall understand later, this problem is intimately related to the construction of a *deterministic* optimal coupling.

The second direction of research came from the work of Yann Brenier. While studying problems in incompressible fluid mechanics, Brenier needed to construct an operator that would act like the projection on the set of measure-preserving mappings in an open set (in probabilistic language, measure-preserving mappings are deterministic couplings of the Lebesgue measure with itself). He understood that he could do so by introducing an optimal coupling: If u is the map of which one wants to compute the projection, introduce a coupling of the Lebesgue measure $\mathcal{L}$ with $u_{\#}\mathcal{L}$. This study revealed an unexpected link between optimal transport and fluid mechanics, and at the same time attracted the attention of the community of partial differential equations, by pointing out its relation with the theory of the Monge-Ampère equation.

The third direction of research, certainly the most surprising, came from outside mathematics. Mike Cullen was part of a group of meteorologists with a well-developed mathematical taste, working on *semi-geostrophic equations*, used in meteorology for the modelling of atmospheric fronts. Cullen and his collaborators showed that a certain well-known change of unknown due to Hoskins could be re-interpreted in terms of an optimal coupling problem, and they identified the minimization property as a *stability* condition. A striking feature of this work was that optimal transport could arise naturally in partial differential equations which seemed to have nothing to do with it.

All three contributions emphasized (in their respective domain) that *important information can be gained by a qualitative description of optimal transport*. These new directions of research attracted various mathematicians (among the first, Luis Caffarelli, Craig Evans,

Wilfrid Gangbo, Robert McCann, and others), who worked on a better description of the structure of optimal transport and found some other applications.

An important conceptual step was accomplished by Felix Otto, who discovered an appealing formalism which in effect introduced a *differential* point of view in optimal transport theory. This opened the way to a more geometric description of the space of probability measures, and connected optimal transport to the theory of diffusion equations, thus leading to a rich interplay between geometry, functional analysis and partial differential equations.

Nowadays optimal transport has become a thriving industry, involving many researchers and many trends. Apart from meteorology, fluid mechanics and diffusion equations, it has also been applied to such diverse topics as the collapse of sandpiles, the matching of images, and the design of networks or reflector antennas. My book, *Topics in Optimal Transportation*, written between 2000 and 2003, was the first attempt to present a synthetic view of the modern theory. Since then the field has grown much faster than I expected, and it was never so active as now.

Bibliographical Notes

Before the twentieth century, the main references about the problem of "déblais et remblais" are the memoirs by Monge [276], Dupin [148] and Appell [22]. Besides achieving important mathematical results, Monge and Dupin were strongly committed to the development of the society at their time and it is interesting to browse some of their writings about economics and industry (a list can be found online at `http://gallica.bnf.fr`). A lively account of Monge's life and political commitments can be found in Bell's delightful treatise, *Men of Mathematics* [**?**].

Kantorovich introduced his minimization problem in [220], established his duality theorem in [221], and later made the link with Monge's problem in [222]. He wrote a short autobiography on the occasion of his Nobel Prize [223].

Mather introduced minimizing measures in [259], and later proved his Lipschitz graph theorem in [260]. The explicit connection with the Monge-Kantorovich problem came only recently [47].

Tanaka's contributions to kinetic theory go back to the mid-seventies [280, 349, 350]. This line of research was later taken back by Toscani and collaborators [60, 305]. These papers constituted my first contact with the optimal transport problem. More recent developments about the use of optimal transport in the Boltzmann equation for granular media appear for instance in [62].

Brenier announced his main results in a short note [69], then published detailed proofs in [72]. Chapter 3 in [365] is entirely devoted to Brenier's *polar factorization theorem* (which includes the existence of the projection operator), its interpretation and consequences. About the sources of inspiration of Brenier, and various links between optimal transport and hydrodynamics, one may consult [70, 74, 75].

The semi-geostrophic system was introduced by Eliassen [152] and Hoskins [213, 214, 215]; it is *very* briefly described in [365, Problem 9, pp. 323–326]; a short list of references is also provided there. Cullen and collaborators wrote many papers on the subject, see in particular [125]; see also the review article [124], the work by Cullen and Gangbo [123], or the recent book by Cullen [121].

Further links between optimal transport and other fields of mathematics (or physics) can be found in my book [365], or in the treatise by Rachev and Rüschendorf [306]. Here below is a non-exhaustive list of some unexpected applications. Relations with the

modelling of sandpiles are reviewed by Evans [154]. Applications of optimal transport to image processing are discussed by Gangbo and McCann [**?**], Haker, Zhu, Tannenbaum and Angenent [**?**], and others. X.-J. Wang [**?**] discovered that the theoretical problem of designing reflector antennas could be recast as an optimal transport problem. In his PhD, Bernot [**?**] made the link between optimal transport and the design of networks.

Many generalizations of optimal transport have been considered, such as the transshipment problem [306], or optimal coupling with more than two given marginals [230, 319, 323, 325, 180]. Other variants are discussed also in [306].

Part I

Qualitative description of optimal transport

The first part of these notes is devoted to the description and characterization of optimal transport under certain regularity assumptions on the measures and the cost function.

As a start, some general theorems about optimal transport plans are established in Chapters 4 and 5, in particular the Kantorovich duality theorem. The emphasis is on c-cyclically monotone maps, both in the statements and in the proofs. The assumptions on the cost function and the spaces will be very general.

From the Monge–Kantorovich problem one can derive natural distance functions on spaces of probability measures, by choosing the cost function as a power of the distance. The main properties of these distances are established in Chapter 6.

In Chapter 7 a time-dependent version of the Monge–Kantorovich problem is investigated, which leads to an interpolation procedure between probability measures, called displacement interpolation. The natural assumption is that the cost function derives from a Lagrangian action, in the sense of classical mechanics; still (almost) no smoothness is required at that level. In Chapter 8 I shall make further assumptions of smoothness and convexity, and recover some regularity properties of the displacement interpolant by a strategy due to Mather.

Then in Chapters 9 and 10 it is shown how to establish the existence of deterministic optimal couplings, and characterize the associated transport maps, again under adequate regularity and convexity assumptions. These transports might not be smooth, as explained in Chapter 12, but the Change of Variables Formula still applies, as discussed in Chapter 11.

The main results of this part are synthetized and summarized in Chapter 13.

4

Basic properties

Existence

The first good thing about optimal couplings is that they exist:

Theorem 4.1 (Existence of an optimal coupling). *Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two Polish (i.e. metric, complete separable) probability spaces. Let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ be a lower semi-continuous cost function, bounded below by some real number. Then there is always a coupling of μ and ν which minimizes the total cost $\mathbb{E}\, c(X, Y)$ among all possible couplings (X, Y).*

The proof relies on basic variational arguments involving the topology of weak convergence (i.e. imposed by bounded continuous test functions). There are two key properties: (a) lower semi-continuity and (b) compactness. These issues are taken care of respectively in Lemmas 4.2 and 4.3 below, which will be used again in the sequel. Before going on, I recall **Prokhorov's theorem**: *If $\mathcal{X}$ is a Polish space, then a set $\mathcal{P} \subset P(\mathcal{X})$ is precompact for the weak topology if and only if it is* tight, *i.e. for any $\varepsilon > 0$ there is a compact set K_ε such that $\mu[\mathcal{X} \setminus K_\varepsilon] \leq \varepsilon$ for all $\mu \in \mathcal{P}$.*

Lemma 4.2 (lower semi-continuity of the cost functional). *Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces, and let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ be a lower semi-continuous cost function, bounded below. Then $F : \pi \to \int c\, d\pi$ is also a lower semi-continuous functional on $P(\mathcal{X} \times \mathcal{Y})$, bounded below.*

Proof of Lemma 4.2. Let π_k be a family of probability measures on $\mathcal{X} \times \mathcal{Y}$, converging weakly to some probability measure π. Since c is lower semi-continuous, it can be written as the pointwise limit of a nondecreasing family $(c_\ell)_{\ell \in \mathbb{N}}$ of continuous real-valued functions. Then

$$\int c\, d\pi = \lim_{\ell \to \infty} \int c_\ell\, d\pi = \lim_{\ell \to \infty} \lim_{k \to \infty} \int c_\ell\, d\pi_k \leq \liminf_{k \to \infty} \int c\, d\pi_k.$$

So the function $F : \pi \longmapsto \int c\, d\pi$ is indeed lower semi-continuous with respect to weak convergence. It is also clearly bounded below by $\inf c$. □

Lemma 4.3 (tightness of transference plans). *Let $\mathcal{X}$ and $\mathcal{Y}$ be two Polish spaces. Let $\mathcal{P} \subset P(\mathcal{X})$ and $\mathcal{Q} \subset P(\mathcal{Y})$ be tight subsets of $P(\mathcal{X})$ and $P(\mathcal{Y})$ respectively. Then the set $\Pi(\mathcal{P}, \mathcal{Q})$ of all transference plans whose marginals lie in $\mathcal{P}$ and $\mathcal{Q}$ respectively, is itself tight in $P(\mathcal{X} \times \mathcal{Y})$.*

Proof of Lemma 4.3. Let $\mu \in \mathcal{P}$, $\nu \in \mathcal{Q}$, and $\pi \in \Pi(\mu, \nu)$. By assumption, for any $\varepsilon > 0$ there is a compact set $K_\varepsilon \subset \mathcal{X}$, independent of the choice of μ in $\mathcal{P}$, such that $\mu[\mathcal{X} \setminus K_\varepsilon] \leq \varepsilon$;

and similarly there is a compact set $L_\varepsilon \subset \mathcal{Y}$, independent of the choice of ν in $\mathcal{Q}$, such that $\nu[\mathcal{Y} \setminus L_\varepsilon] \le \varepsilon$. It follows that, for any coupling (X, Y) of (μ, ν),

$$\mathbb{P}\left[(X,Y) \notin K_\varepsilon \times L_\varepsilon\right] \le \mathbb{P}\left[X \notin K_\varepsilon\right] + \mathbb{P}\left[Y \notin L_\varepsilon\right] \le 2\varepsilon.$$

The desired result follows since this bound is independent of the coupling, and $K_\varepsilon \times L_\varepsilon$ is compact in $\mathcal{X} \times \mathcal{Y}$. □

Proof of Theorem 4.1. Since $\mathcal{X}$ is Polish, $\{\mu\}$ is tight in $P(\mathcal{X})$; and similarly, $\{\nu\}$ is tight in $P(\mathcal{Y})$. So $\Pi(\mu, \nu)$ is tight in $P(\mathcal{X} \times \mathcal{Y})$, and by Prokhorov's theorem this set has a compact closure. By passing to the limit in the equation for marginals, we see that $\Pi(\mu, \nu)$ is closed, so it is in fact *compact*.

Since $F : \pi \longmapsto \int c\, d\pi$ is a lower semi-continuous function on a compact set, bounded below, it admits a minimizer (this is a standard result; to prove it, select a minimizing sequence and pass to the lim inf). So there is indeed a transference plan π which minimizes the cost functional $\int c\, d\pi$. □

Remark 4.4. This existence theorem does not imply that the optimal cost is finite. It might be that *all* transport plans lead to an infinite total cost. A simple condition to rule out this annoying possibility is

$$\int c(x,y)\, d\mu(x)\, d\nu(y) < +\infty;$$

which guarantees that at least the independent coupling has finite total cost. In the sequel, I shall sometimes make the stronger assumption

$$c(x,y) \le c_{\mathcal{X}}(x) + c_{\mathcal{Y}}(y), \qquad (c_{\mathcal{X}}, c_{\mathcal{Y}}) \in L^1(d\mu) \times L^1(d\nu),$$

which implies that *any* coupling has finite total cost, and has other nice consequences (see e.g. Theorem 5.9).

Restriction property

The second good thing about optimal couplings is that *any sub-coupling is still optimal*. In words: If you have an optimal transport plan, then any induced sub-plan (transferring part of the initial mass to part of the final mass) has to be optimal too — otherwise you might be able to lower the cost of the sub-plan, and as a consequence the cost of the whole plan. This is the content of the next theorem.

Theorem 4.5 (Optimality is inherited by restriction). *With the same notation as in Theorem 4.1, let π be the law of an optimal coupling between two probability measures μ and ν. Let $\widetilde{\pi}$ be a nonnegative measure on $\mathcal{X} \times \mathcal{Y}$, such that $\widetilde{\pi} \le \pi$ and $\widetilde{\pi}[\mathcal{X} \times \mathcal{Y}] > 0$. Then the probability measure*

$$\pi' := \frac{\widetilde{\pi}}{\widetilde{\pi}[\mathcal{X} \times \mathcal{Y}]}$$

is an optimal transference plan between its marginals μ' and ν'.

Moreover, if π is the unique optimal transference plan between μ and ν, then also π' is the unique optimal transference plan between μ' and ν'.

Example 4.6. If (X, Y) is an optimal coupling of (μ, ν), and $\mathcal{Z} \subset \mathcal{X} \times \mathcal{Y}$ is such that $\mathbb{P}\big[(X, Y) \in \mathcal{Z}\big] > 0$, then the couple (X, Y), conditioned to lie in $\mathcal{Z}$, is an optimal coupling of (μ', ν'), where μ' (resp. ν') is the law of X (resp. Y) conditioned by the event "$(X, Y) \in \mathcal{Z}$".

Proof of Theorem 4.5. Assume that π' is not optimal; then there exists π'' such that

$$(\mathrm{proj}_{\mathcal{X}})_{\#}\pi'' = (\mathrm{proj}_{\mathcal{X}})_{\#}\pi' =: \mu', \qquad (\mathrm{proj}_{\mathcal{Y}})_{\#}\pi'' = (\mathrm{proj}_{\mathcal{Y}})_{\#}\pi' =: \nu', \tag{4.1}$$

yet

$$\int c(x, y)\, d\pi''(x, y) < \int c(x, y)\, d\pi'(x, y). \tag{4.2}$$

Then consider

$$\widehat{\pi} := (\pi - \widetilde{\pi}) + \widetilde{Z}\pi'', \tag{4.3}$$

where $\widetilde{Z} = \widetilde{\pi}[\mathcal{X} \times \mathcal{Y}] > 0$. Clearly, $\widehat{\pi}$ is a nonnegative measure. On the other hand, it can be written as

$$\widehat{\pi} = \pi + \widetilde{Z}(\pi'' - \pi');$$

then (4.1) shows that $\widehat{\pi}$ has the same marginals as π, while (4.2) implies that it has a lower transport cost than π. This contradicts the optimality of π. The conclusion is that π' is in fact optimal.

It remains to prove the last statement of Theorem 4.5. Assume that π is the unique optimal transference plan between μ and ν; and let π'' be any optimal transference plan between μ' and ν'. Define again $\widehat{\pi}$ by (4.3). Then $\widehat{\pi}$ has the same cost as π, so $\widehat{\pi} = \pi$, which implies that $\widetilde{\pi} = \widetilde{Z}\pi''$, i.e. $\pi'' = \pi'$. □

Convexity properties

The following estimates are of constant use:

Theorem 4.7 (Convexity of the optimal cost). *Let $\mathcal{X}$ and $\mathcal{Y}$ be two Polish space, let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ be a lower semi-continuous function, and let C be the associated optimal transport cost functional on $P(\mathcal{X}) \times P(\mathcal{Y})$. Let (Θ, λ) be a probability space, and let μ_θ, ν_θ be two measurable functions defined on Θ, with values in $P(\mathcal{X})$ and $P(\mathcal{Y})$ respectively. Then*

$$C\left(\int_\Theta \mu_\theta\, \lambda(d\theta),\ \int_\Theta \nu_\theta\, \lambda(d\theta)\right) \ \le\ \left(\int_\Theta C(\mu_\theta, \nu_\theta)\, \lambda(d\theta)\right)^{\frac{1}{p}}.$$

Proof. For each θ, let π_θ be optimal in the transport problem with cost d^p and marginals μ_θ and ν_θ. Then $\pi := \int \pi_\theta\, \lambda(d\theta)$ has marginals $\mu := \int \mu_\theta\, \lambda(d\theta)$ and $\nu := \int \nu_\theta\, \lambda(d\theta)$. So

$$\begin{aligned} C(\mu, \nu) &\le \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, d\pi(x, y) \\ &= \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\ \left(\int_\Theta \pi_\theta\, \lambda(d\theta)\right)(dx\, dy) \\ &= \int_\Theta \left(\int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, \pi_\theta(dx\, dy)\right)\ \lambda(d\theta) \\ &= \int_\Theta C(\mu_\theta, \nu_\theta)\, \lambda(d\theta), \end{aligned}$$

and then the conclusion follows. □

Description of optimal plans

Obtaining more precise information about minimizers will be much more of a sport. Here is a short list of questions that one might ask:

- is the optimal coupling unique? smooth in some sense?

- is there a *Monge coupling*, i.e. a deterministic optimal coupling?

- is there a geometrical way to characterize optimal couplings? Can one check in practice that a certain coupling is optimal?

About the second question: Why don't we try to apply the same reasoning as in the proof of Theorem 4.1? The problem is that the set of deterministic couplings is in general *not* compact. In fact, this set is often dense in the larger space of all couplings! So we may expect that the *value* of the infimum in the Monge problem coincides with the value of the minimum in the Kantorovich problem; but there is no a priori reason to expect the existence of a Monge minimizer.

Example 4.8. Let $\mathcal{X} = \mathcal{Y} = \mathbb{R}^2$, let $c(x, y) = |x-y|^2$, let μ be $\mathcal{H}^1$ restricted to $\{0\}\times[-1, 1]$, and let ν be $(1/2)\mathcal{H}^1$ restricted to $\{-1, 1\} \times [-1, 1]$, where $\mathcal{H}^1$ is the one-dimensional Hausdorff measure. Then there is a unique optimal transport, which for each point $(0, a)$ sends one half of the mass at $(0, a)$ to $(-1, a)$, and the other half to $(1, a)$. This is not a Monge transport, but it is easy to approximate it by Monge transports.

Fig. 4.1. The optimal plan, represented on the left image, consists in splitting the mass in the center into two halves and transport mass horizontally. On the right the filled regions represent the lines of transport for a deterministic (without splitting of mass) approximation of the optimum.

Bibliographical Notes

Theorem 4.1 has probably been known from immemorial times. Prokhorov's theorem is a most classical result that can be found e.g. in [52] (or in my own course on integration, accessible online via `http://www.umpa.ens-lyon.fr/~cvillani/Cours`).

Theorems of the form "infimum cost in the Monge problem = minimum cost in the Kantorovich problem" have been established by Gangbo [177, Appendix A], Ambrosio [10, Theorem 2.1] and Pratelli [301, Theorem B]. The most general results to this date are those which appear in Pratelli's work: The equality holds true if the source space $(\mathcal{X}, \mu)$ is Polish without atoms, and the cost is continuous $\mathcal{X}\times\mathcal{Y} \to \mathbb{R}\cup\{+\infty\}$ (with the value $+\infty$ allowed).

5

Cyclical monotonicity and Kantorovich duality

To get started, we should first become acquainted with two basic concepts in the theory of optimal transport. The first one is a geometric property called *cyclical monotonicity*; the second one is the *Kantorovich dual problem*, which is another face of the original Monge–Kantorovich problem.

Definitions and heuristics

I shall start by explaining the concepts of cyclical monotonicity and Kantorovich duality in an informal way, sticking to the bakery analogy. Assume you have been hired by a large consortium of bakeries and cafés, to be in charge of the distribution of bread from production units (bakeries) to consumption units (cafés). The locations of the bakeries and cafés, their respective production and consumption rates, are all determined in advance. You have written a transference plan, which says, for each bakery (located at) x_i and each café y_j, how much bread should go each morning from x_i to y_j.

As there are complaints that the transport cost associated with your plan is actually too high, you try to reduce it. For that purpose you choose a bakery x_1 that sends part of its production to a distant café y_1, and decide that one basket of bread will be rerouted to another café y_2, that is closer to x_1; thus you will gain $c(x_1, y_2) - c(x_1, y_1)$. Of course, now this results in an excess of bread in y_2, so one basket of bread arriving to y_2 (say, from bakery x_2) should in turn be rerouted to yet another café, say y_3. The process goes on and on until finally you redirect a basket from some bakery x_N to y_1, at which point you can stop since you have a new admissible transference plan.

The new plan is (strictly) better than the previous one if and only if

$$c(x_1, y_2) + c(x_2, y_3) + \ldots + c(x_N, y_1) < c(x_1, y_1) + c(x_2, y_2) + \ldots + c(x_N, y_N).$$

Thus, if you can find such *cycles* $(x_1, y_1), \ldots, (x_N, y_N)$ in your transference plan, certainly the latter is not optimal. On the contrary, if you do *not* find then, then your plan cannot be improved (at least by the procedure described above) and it is likely to be optimal. This motivates the following definitions.

Definition 5.1 (cyclical monotonicity). *Let $\mathcal{X}, \mathcal{Y}$ be arbitrary sets, and $c : \mathcal{X} \times \mathcal{Y} \to (-\infty, +\infty]$ be a cost function. A subset $\Gamma \subset \mathcal{X} \times \mathcal{Y}$ is said to be c-cyclically monotone if, for any $N \in \mathbb{N}$, and any family $(x_1, y_1), \ldots, (x_N, y_N)$ of points in Γ, holds the inequality*

$$\sum_{i=1}^{N} c(x_i, y_i) \leq \sum_{i=1}^{N} c(x_i, y_{i+1}) \tag{5.1}$$

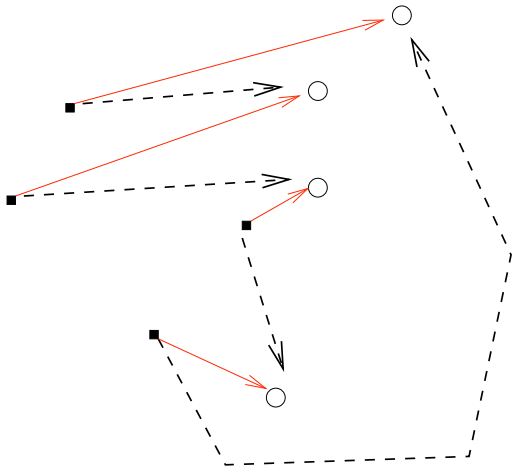


Fig. 5.1. An attempt to improve the cost by a cycle; solid arrows indicate the mass transport in the original plan, dashed arrows the paths along which a bit of mass is rerouted.

(with the convention $y_{N+1} = y_1$). A transference plan is said to be c-cyclically monotone if it is concentrated on a c-cyclically monotone set.

Informally, a c-cyclically monotone plan is a plan that *cannot be improved*: it is impossible to perturb it (in the sense considered before, by rerouting mass along some cycle) and get something more economical. One can think of it as a kind of local minimizer. It is intuitively obvious that an optimal plan should be c-cyclically monotone; the converse property is much less obvious (maybe it is possible to get something better by radically changing the plan), but we shall soon see that it holds true under mild conditions.

The next key concept is the dual Kantorovich problem. While the central notion in the original Monge–Kantorovich problem is *cost*, in the dual problem it is *price*. Imagine that a company offers to take care of all your transportation problem, buying bread at the bakeries and selling them to the cafés; what happens in between is not your problem (and maybe they have tricks to do the transport at a lower price than you). Let $\psi(x)$ be the price at which a basket of bread is bought at bakery x, and $\phi(y)$ the price at which it is sold at café y. On the whole, the price which the consortium bakery+café pays for the transport is $\phi(y) - \psi(x)$, instead of the original cost $c(x, y)$. This of course is for each unit of bread: if there is a mass $\mu(dx)$ at x, then the total price of the bread shipment from there will be $\psi(x)\,\mu(dx)$.

So as to be competitive, the company needs to set up prices in such a way that

$$\forall (x, y), \qquad \phi(y) - \psi(x) \le c(x, y). \tag{5.2}$$

When you were handling the transportation yourself, your problem was to *minimize the cost*. Now that the company takes this into charge, their problem is to *maximize the profits*. This naturally leads to the **dual Kantorovich problem**:

$$\sup \left\{ \int_{\mathcal{Y}} \phi(y)\, d\nu(y) - \int_{\mathcal{X}} \psi(x)\, d\mu(x); \quad \phi(y) - \psi(x) \le c(x, y) \right\}. \tag{5.3}$$

From a mathematical point of view, it will be imposed that the functions (ψ, ϕ) appearing in (5.3) be integrable: $\psi \in L^1(\mathcal{X}, \mu)$; $\phi \in L^1(\mathcal{Y}, \nu)$.

With the intervention of the company, the shipment of each unit of bread does not cost more than it used to when you were handling it yourself; so it is obvious that the supremum in (5.3) is less than the optimal transport cost:

$$\sup_{\phi-\psi\leq c}\left\{\int_{\mathcal{Y}}\phi(y)\,d\nu(y)-\int_{\mathcal{X}}\psi(x)\,d\mu(x)\right\}\leq\inf_{\pi\in\Pi(\mu,\nu)}\left\{\int_{\mathcal{X}\times\mathcal{Y}}c(x,y)\,d\pi(x,y)\right\}. \tag{5.4}$$

Clearly, if we can find a pair (ψ,ϕ) and a transference plan π for which there is equality, then (ψ,ϕ) is optimal in the left-hand side and π is also optimal in the right-hand side.

A pair of price functions (ψ,ϕ) will informally be said to be *competitive* if it satisfies (5.2). For a given y, it is of course in the interest of the company to set the highest possible competitive price $\phi(y)$, i.e. the highest lower bound (i.e. the infimum) for $\psi(x)+c(x,y)$, among all bakeries x. Similarly, for a given x, the price $\psi(x)$ should be the supremum of all $\phi(y)-c(x,y)$. Thus it makes sense to describe a pair of prices (ψ,ϕ) as *tight* if

$$\phi(y)=\inf_x\Big(\psi(x)+c(x,y)\Big),\qquad \psi(x)=\sup_y\Big(\phi(y)-c(x,y)\Big). \tag{5.5}$$

In words, prices are tight if it is impossible for the company to raise the selling price, or lower the buying price, without losing its competitivity.

Consider an arbitrary pair of competitive prices (ψ,ϕ). We can always improve ϕ by replacing it by $\phi_1(y)=\inf_x\big(\psi(x)+c(x,y)\big)$. Then we can also improve ψ by replacing by $\psi_1(x)=\sup_y\big(\phi_1(y)-c(x,y)\big)$; then replacing ϕ_1 by $\phi_2(y)=\inf_x\big(\psi_1(x)+c(x,y)\big)$, and so on. It turns out that this process is stationary: as an easy exercise, the reader can check that $\phi_2=\phi_1$, $\psi_2=\psi_1$, which means that after just one iteration we have obtained a pair of tight prices. Thus, when we consider the dual Kantorovich problem (5.3), it makes sense to restrict our attention to tight pairs, in the sense of equation (5.5). From that equation we can reconstruct ϕ in terms of ψ, so we can just take ψ as the only unknown in our problem. That unknown cannot be just any function: if you take a general function ψ, compute ϕ by the first formula in (5.5), there is no chance that the second formula will be satisfied. In fact this second formula will hold true if and only if ψ is *c-convex*, in the sense of the following definition.

Definition 5.2 (c-convexity). *Let $\mathcal{X},\mathcal{Y}$ be sets, and $c:\mathcal{X}\times\mathcal{Y}\to(-\infty,+\infty]$ be a fixed cost. A function $\psi:\mathcal{X}\to\mathbb{R}\cup\{+\infty\}$ is said to be c-convex if it is not identically $+\infty$, and there exists $\zeta:\mathcal{Y}\to\mathbb{R}\cup\{\pm\infty\}$ such that*

$$\forall x\qquad \psi(x)=\sup_y\Big(\zeta(y)-c(x,y)\Big). \tag{5.6}$$

Then its c-transform is the function ψ^c defined by

$$\forall y\qquad \psi^c(y)=\inf_x\Big(\psi(x)+c(x,y)\Big), \tag{5.7}$$

and its c-subdifferential is the c-cyclically monotone set defined by

$$\partial_c\psi:=\Big\{(x,y)\subset\mathcal{X}\times\mathcal{Y};\quad \psi^c(y)-\psi(x)=c(x,y)\Big\}.$$

The functions ψ and ψ^c are said to be c-conjugate.

Then the c-subdifferential of ψ at point x is

$$\partial_c\psi(x)=\Big\{y\in\mathcal{Y};\quad (x,y)\in\partial_c\psi\Big\},$$

or equivalently

$$\forall z\in\mathcal{X},\qquad \psi(x)+c(x,y)\leq\psi(z)+c(z,y). \tag{5.8}$$

Particular Case 5.3. If $c(x,y) = -x \cdot y$ on $\mathbb{R}^n \times \mathbb{R}^n$, then then the c-transform coincides with the usual Legendre transform, and c-convexity is just plain convexity on $\mathbb{R}^n$. (Actually, this is a slight oversimplification: c-convexity is equivalent to plain convexity plus lower semi-continuity! A convex function is automatically continuous on the largest open set Ω where it is finite, but lower semi-continuity might fail at the boundary of Ω...) One can think of the cost function $c(x,y) = -x \cdot y$ as basically the same as $c(x,y) = |x-y|^2/2$, since the "interaction" between the positions x and y is the same for both costs.

Particular Case 5.4. If $c = d$ is a *distance* on some metric space $\mathcal{X}$, or more generally if c satisfies the triangular inequality $c(x,z) \le c(x,y) + c(y,z)$, then a c-convex function is just a 1-Lipschitz function, and it is its own c-transform. Indeed, if ψ is c-convex then it is obviously 1-Lipschitz; conversely, if ψ is 1-Lipschitz, then $\psi(x) \le \psi(y) + d(x,y)$, so $\psi(x) = \inf_y [\psi(y) + d(x,y)] = \psi^c(x)$. As an even more particular case, if $c(x,y) = 1_{x \neq y}$, then ψ is c-convex if and only if $\sup \psi - \inf \psi \le 1$, and then again $\psi^c = \psi$.

Remark 5.5. There is no measure theory in Definition 5.2, so no assumption of measurability is made, and the supremum in (5.6) is a true supremum, not just an essential supremum; the same is true for the infimum in (5.7). If c is continuous, then a c-convex function is automatically lower semi-continuous, and its subdifferential is closed; but if c is not continuous the measurability of ψ and $\partial_c \psi$ are not guaranteed.

Remark 5.6. I excluded the case when $\psi \equiv +\infty$ so as to avoid trivial situations; what is called a c-convex function here might more properly (!) be called a *proper* c-convex function. This automatically implies that ζ in (5.6) does not take the value $+\infty$ at all if c is real-valued. If c does achieve infinite values, then the correct convention in (5.6) is $(+\infty) - (+\infty) = -\infty$.

If ψ is a function on $\mathcal{X}$, then its c-transform is a function on $\mathcal{Y}$. Conversely, given a function on $\mathcal{Y}$, one may define its c-transform as a function on $\mathcal{X}$. It will be convenient in the sequel to define the latter concept by an *infimum* rather than a supremum. This convention has the drawback to break the symmetry between the roles of $\mathcal{X}$ and $\mathcal{Y}$, but it has other advantages that will be apparent later on.

Definition 5.7 (c-concavity). *With the same notation as in Definition 5.2, a function $\phi : \mathcal{Y} \to \mathbb{R} \cup \{-\infty\}$ is said to be c-concave if it is not identically $-\infty$, and there exists $\psi : \mathcal{X} \to \mathbb{R} \cup \{\pm\infty\}$ such that $\phi = \psi^c$.*

Then its c-transform is the function ϕ^c defined by

$$\forall x \qquad \phi^c(x) = \sup_y \Big(\phi(y) - c(x,y) \Big);$$

and its c-superdifferential is the c-cyclically monotone set defined by

$$\partial_c \psi := \Big\{ (x,y) \subset \mathcal{X} \times \mathcal{Y}; \quad \psi^c(y) - \psi(x) = c(x,y) \Big\}.$$

The following proposition may be taken as the main justification for the concept of c-convexity.

Proposition 5.8 (Alternative characterization of c-convexity via c-convexification). *For any function $\psi : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$, define its c-convexification to be $\psi^{cc} = (\psi^c)^c$. More explicitly,*

$$\psi^{cc}(x) = \sup_{y \in \mathcal{Y}} \inf_{\widetilde{x} \in \mathcal{X}} \Big(\psi(\widetilde{x}) + c(\widetilde{x}, y) - c(x,y) \Big).$$

Then ψ is c-convex if and only if $\psi^{cc} = \psi$.

Proof of Proposition 5.8. As a general fact, for any function $\phi : \mathcal{Y} \to \mathbb{R} \cup \{-\infty\}$ (not necessarily c-convex), one has the identity $\phi^{ccc} = \phi^c$. Indeed,

$$\phi^{ccc}(x) = \sup_y \inf_{\widetilde{x}} \sup_{\widetilde{y}} \Big[\phi(\widetilde{y}) - c(\widetilde{x}, \widetilde{y}) + c(\widetilde{x}, y) - c(x, y) \Big];$$

then the choice $x = \widetilde{x}$ shows that $\phi^{ccc}(x) \le \phi^c(x)$; while the choice $\widetilde{y} = y$ shows that $\phi^{ccc}(x) \ge \phi^c(x)$.

So if ψ is c-convex, then there is ζ such that $\psi = \zeta^c$, so $\psi^{cc} = \zeta^{ccc} = \zeta^c = \psi$.

The converse is obvious: If $\psi^{cc} = \psi$, then ψ is the c-transform of ψ^c, so it is c-convex. □

Kantorovich duality

We are now ready to state and prove the main result in this chapter.

Theorem 5.9 (Kantorovich duality). *Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two Polish probability spaces and let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ be a lower semi-continuous cost function, such that*

$$\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, \qquad c(x, y) \ge a(x) + b(y)$$

for some real-valued, upper semi-continuous functions $a \in L^1(\mu)$, $b \in L^1(\nu)$. Then

(i) There is duality:

$$\begin{aligned}
\min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\, d\pi(x,y) &= \sup_{(\psi,\phi) \in C_b(\mathcal{X}) \times C_b(\mathcal{Y});\ \phi - \psi \le c} \left(\int_{\mathcal{Y}} \phi(y)\, d\nu(y) - \int_{\mathcal{X}} \psi(x)\, d\mu(x) \right) \\
&= \sup_{(\psi,\phi) \in L^1(\mu) \times L^1(\nu);\ \phi - \psi \le c} \left(\int_{\mathcal{Y}} \phi(y)\, d\nu(y) - \int_{\mathcal{X}} \psi(x)\, d\mu(x) \right) \\
&= \sup_{\psi \in L^1(\mu)} \left(\int_{\mathcal{Y}} \psi^c(y)\, d\nu(y) - \int_{\mathcal{X}} \psi(x)\, d\mu(x) \right) \\
&= \sup_{\phi \in L^1(\nu)} \left(\int_{\mathcal{Y}} \phi(y)\, d\nu(y) - \int_{\mathcal{X}} \phi^c(x)\, d\mu(x) \right),
\end{aligned}$$

and in the above suprema one might as well impose that ψ be c-convex, and ϕ be c-concave.

(ii) If c is real-valued and the optimal cost $C(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int c\, d\pi$ is finite, then there is a measurable c-cyclically monotone set $\Gamma \subset \mathcal{X} \times \mathcal{Y}$ (closed if a, b, c are continuous) such that for any $\pi \in \Pi(\mu, \nu)$ the following five statements are equivalent:

(a) π is optimal;

(b) π is c-cyclically monotone;

(c) There is a c-convex ψ such that, π-almost surely, $\psi^c(y) - \psi(x) = c(x, y)$;

(d) There are functions $\psi : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ and $\phi : \mathcal{Y} \to \mathbb{R} \cup \{-\infty\}$, such that $\phi(y) - \psi(x) \le c(x, y)$ for all (x, y), with equality π-almost surely.

(e) π is concentrated on Γ.

(iii) If c is real-valued, $C(\mu, \nu) < +\infty$, and one has the pointwise upper bound

$$c(x, y) \le c_{\mathcal{X}}(x) + c_{\mathcal{Y}}(y), \qquad (c_{\mathcal{X}}, c_{\mathcal{Y}}) \in L^1(\mathcal{X}, \mu) \times L^1(\mathcal{Y}, \nu), \tag{5.9}$$

then both the original and the dual Kantorovich problems admit solutions, so

$$\min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\, d\pi(x,y) = \max_{(\psi,\phi)\in L^1(\mu)\times L^1(\nu);\ \phi-\psi\le c} \left(\int_{\mathcal{Y}} \phi(y)\, d\nu(y) - \int_{\mathcal{X}} \psi(x)\, d\mu(x) \right) \tag{5.10}$$

$$= \max_{\psi \in L^1(\mu)} \left(\int_{\mathcal{Y}} \psi^c(y)\, d\nu(y) - \int_{\mathcal{X}} \psi(x)\, d\mu(x) \right), \tag{5.11}$$

and in the latter expressions one might as well impose that ψ be c-convex and $\phi = \psi^c$. If in addition a, b and c are continuous, then there is a closed c-cyclically monotone set $\Gamma \subset \mathcal{X} \times \mathcal{Y}$, such that

$$\begin{cases} \pi \in \Pi(\mu,\nu) \text{ is optimal in the Kantorovich problem if and only if } \pi \text{ is concentrated on } \Gamma; \\ \psi \text{ is optimal in the dual Kantorovich problem if and only if } \Gamma \subset \partial_c \psi. \end{cases}$$

Remark 5.10. When the cost c is continuous, then the support of π is c-monotone; but for a discontinuous cost function it might a priori be true that π is concentrated on a (non-closed) monotone set, while the support of π is not monotone. So, in the sequel, the words "concentrated on" are not exchangeable with "supported in". There is another subtlety for discontinuous cost functions: It is not clear that the functions ϕ and ψ^c appearing in statements (ii) and (iii) are Borel measurable; it will only be proven that they coincide with measurable functions outside of a ν-negligible set.

Remark 5.11. Note the difference between statements (b) and (e). The set Γ appearing in (ii)(b) is the same for all optimal π's, it only depends on μ and ν. This set is in general not unique. If c is continuous and Γ is imposed to be closed, then one can define a smallest Γ, which is the closure of the union of all the supports of the optimal π's. There is also a largest Γ, which is the intersection of all the subdifferentials $\partial_c \psi$, where ψ is such that there exists an optimal π supported in $\partial_c \psi$. (Since the cost function is assumed to be continuous, the subdifferentials are closed, and so is their intersection.)

Remark 5.12. Here is a useful practical consequence of Theorem 5.9: Given a transference plan π, if you can cook up a pair of competitive prices (ψ, ϕ) such that $\phi(y) - \psi(x) = c(x,y)$ throughout the support of π, then you know that π is optimal. This theorem also shows that optimal transference plans satisfy very special conditions: if you fix an optimal pair (ψ, ϕ), then mass arriving at y can come from x *only* if $c(x,y) = \phi(y) - \psi(x) = \psi^c(y) - \psi(x)$, which means that

$$x \in \operatorname*{Arg\,min}_{x' \in \mathcal{X}} \Big(\psi(x') + c(x',y) \Big).$$

In terms of my bakery analogy this can be restated as follows: a café accepts bread from a bakery only if the combined cost of buying the bread there and transporting it here is lowest among all possible bakeries. Similarly, given a pair of competitive prices (ψ, ϕ), if we can cook up a transference plan π such that $\phi(y) - \psi(x) = c(x,y)$ throughout the support of π, then you know that (ψ, ϕ) is a solution to the dual Kantorovich problem.

Remark 5.13. The assumption $c \le c_{\mathcal{X}} + c_{\mathcal{Y}}$ in (iii) can be weakened into

$$\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\, d\mu(x)\, d\nu(y) < +\infty,$$

or even

$$\mu \left[\left\{ x;\ \int_{\mathcal{Y}} c(x,y)\, d\nu(y) < +\infty \right\} \right] > 0; \qquad \nu \left[\left\{ y;\ \int_{\mathcal{X}} c(x,y)\, d\mu(x) < +\infty \right\} \right] > 0. \tag{5.12}$$

Particular Case 5.14. The Particular Case 5.4 leads to the following variant of Theorem 5.9. When $c(x,y) = d(x,y)$ is a distance on a Polish space $\mathcal{X}$, and μ, ν belong to $P_1(\mathcal{X})$, then

$$\inf \mathbb{E}\, d(X,Y) = \sup \mathbb{E}\,[\psi(X) - \psi(Y)] = \sup \left\{ \int_{\mathcal{X}} \psi \, d\mu - \int_{\mathcal{Y}} \psi \, d\nu \right\}. \tag{5.13}$$

where the infimum on the left is over all couplings (X,Y) of (μ,ν), and the supremum on the right is over all 1-Lipschitz functions ψ. This is the **Kantorovich–Rubinstein formula**; it holds true as soon as the supremum in the left-hand side is finite, and it is very useful.

Particular Case 5.15. Now consider $c(x,y) = -x \cdot y$ in $\mathbb{R}^n \times \mathbb{R}^n$. This cost is not nonnegative, but we have the lower bound $c(x,y) \geq -(|x|^2 + |y|^2)/2$. So if $x \to |x|^2 \in L^1(\mu)$ and $y \to |y|^2 \in L^1(\nu)$, then one can invoke the Particular Case 5.3 to deduce from Theorem 5.9 that

$$\sup \mathbb{E}\,(X \cdot Y) = \inf \mathbb{E}\left[\varphi(X) + \varphi^*(Y)\right] = \inf \left\{ \int_{\mathcal{X}} \varphi \, d\mu + \int_{\mathcal{Y}} \varphi^* \, d\nu \right\}, \tag{5.14}$$

where the supremum on the left is over all couplings (X,Y) of (μ,ν), the infimum on the right is over all (lower semi-continuous) convex functions on $\mathbb{R}^n$, and φ^* stands for the usual Legendre transform of φ. In formula (5.14), the signs have been changed with respect to the statement of Theorem 5.9, so the problem is to *maximize the correlation* of the random variables X and Y.

Before proving Theorem 5.9, I shall first informally explain the construction. At first reading, the reader might be content with those informal explanations and skip the rigorous proof.

Idea of proof of Theorem 5.9. Take an optimal π (which exists from Theorem 4.1), and let (ψ, ϕ) be two competitive prices. Of course, as in (5.4),

$$\int c(x,y)\, d\pi(x,y) \leq \int \phi \, d\nu - \int \psi \, d\mu = \int [\phi(y) - \psi(x)]\, d\pi(x,y).$$

So if both quantities are equal, then necessarily

$$\phi(y) - \psi(x) = c(x,y) \qquad \pi - \text{almost surely.}$$

Intuitively speaking, wherever there is some transfer of goods from x to y, then the prices should be adjusted exactly to the transport cost.

Now let $(x_i)_{0 \leq i \leq m}$ and $(y_i)_{0 \leq i \leq m}$ be such that (x_i, y_i) belongs to the support of π, so there is indeed some transfer from x_i to y_i. Then we hope that

$$\begin{cases} \phi(y_0) - \psi(x_0) = c(x_0, y_0) \\ \phi(y_1) - \psi(x_1) = c(x_1, y_1) \\ \dots \\ \phi(y_m) - \psi(x_m) = c(x_m, y_m). \end{cases}$$

On the other hand, if x is an arbitrary point,

$$\begin{cases} \phi(y_0) - \psi(x_1) \le c(x_1, y_0) \\ \phi(y_1) - \psi(x_2) \le c(x_2, y_1) \\ \dots \\ \phi(y_m) - \psi(x) \le c(x, y_m). \end{cases}$$

By subtracting these inequalities from the previous equalities, and adding up everything, one obtains

$$\psi(x) \ge \psi(x_0) + \big[c(x_0, y_0) - c(x_1, y_0)\big] + \ldots + \big[c(x_m, y_m) - c(x, y_m)\big].$$

Of course, one can add an arbitrary constant to ψ, provided that one subtract the same constant from ϕ; so it is possible to decide that $\psi(x_0) = 0$, where (x_0, y_0) is arbitrarily chosen in the support of π. Then

$$\psi(x) \ge \big[c(x_0, y_0) - c(x_1, y_0)\big] + \ldots + \big[c(x_m, y_m) - c(x, y_m)\big], \tag{5.15}$$

and this should be true for all choices of (x_i, y_i) $(1 \le i \le m)$ in the support of π, and for all $m \ge 1$. So it becomes natural to *define* ψ as the supremum of all the functions (of the variable x) appearing in the right-hand side of (5.15). It will turn out that this ψ satisfies the equation

$$\psi^c(y) - \psi(x) = c(x, y) \qquad \pi(dx\, dy)\text{-almost surely.}$$

Then, if ψ and ψ^c are integrable, one can write

$$\int c\, d\pi = \int \psi^c\, d\pi - \int \psi\, d\pi = \int \psi^c\, d\nu - \int \psi\, d\mu.$$

This shows at the same time that π is optimal in the original Kantorovich problem, and that the pair (ψ, ψ^c) is optimal in the dual Kantorovich problem. □

Rigorous proof of Theorem 5.9, Part (i). The argument will be divided into a number of steps.

Step 0: *It is sufficient to treat the case when c is nonnegative.*

Indeed, let

$$\widetilde{c}(x, y) := c(x, y) - a(x) - b(y) \ge 0, \qquad \Lambda := \int a\, d\mu + \int b\, d\nu \quad \in \mathbb{R}.$$

Whenever $\psi : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ and $\phi : \mathcal{Y} \to \mathbb{R} \cup \{-\infty\}$ are two functions, define

$$\widetilde{\psi}(x) := \psi(x) + a(x), \qquad \widetilde{\phi}(y) := \phi(y) - b(y).$$

Then the following properties are readily checked:

$$c \text{ lower semi-continuous} \quad \Longrightarrow \quad \widetilde{c} \text{ lower semi-continuous}$$

$$\widetilde{\psi} \in L^1(\mu) \Longleftrightarrow \psi \in L^1(\mu); \qquad \widetilde{\phi} \in L^1(\nu) \Longleftrightarrow \phi \in L^1(\nu);$$

$$\forall \pi \in \Pi(\mu, \nu), \qquad \int \widetilde{c}\, d\pi = \int c\, d\pi + \Lambda;$$

$$\forall (\psi, \phi) \in L^1(\mu) \times L^1(\nu), \quad \int \widetilde{\phi}\, d\nu - \int \widetilde{\psi}\, d\mu = \int \phi\, d\nu - \int \psi\, d\nu \ + \Lambda;$$

$$\psi \text{ is } c\text{-convex} \Longleftrightarrow \widetilde{\psi} \text{ is } \widetilde{c}\text{-convex}; \qquad \phi \text{ is } c\text{-concave} \Longleftrightarrow \widetilde{\phi} \text{ is } \widetilde{c}\text{-concave};$$

(ϕ, ψ) is a pair of c-conjugate functions $\Longleftrightarrow$ $(\widetilde{\phi}, \widetilde{\psi})$ is a pair of $\widetilde{c}$-conjugate functions;

$$\forall \Gamma \subset \mathcal{X} \times \mathcal{Y}, \qquad \Gamma \text{ is } c\text{-cyclically monotone} \Longleftrightarrow \Gamma \text{ is } \widetilde{c}\text{-cyclically monotone}.$$

From these formulas it follows that it is equivalent to establish Theorem 5.9 for the cost c or for the nonnegative cost $\widetilde{c}$. So in the sequel, I shall always assume, without further comment, that c is nonnegative.

Step 1: *If $\mu = (1/n) \sum \delta_{x_i}$, $\nu = (1/n) \sum \delta_{y_j}$, then there is at least one cyclically monotone transference plan.*

Indeed, in that particular case, a transference plan between μ and ν can be identified with a bistochastic array of $n \times n$ real numbers $a_{ij} \in [0, 1]$: each a_{ij} tells what proportion of the $1/n$ mass carried by point x_i will go to destination y_j. So the Monge–Kantorovich problem becomes

$$\inf_{(a_{ij})} \sum_{ij} a_{ij}\, c(x_i, y_i)$$

where the infimum is over all arrays (a_{ij}) satisfying

$$\sum_i a_{ij} = 1, \qquad \sum_j a_{ij} = 1. \tag{5.16}$$

Here we are minimizing a linear function on the compact set $[0, 1]^{n \times n}$, so obviously there exists a minimizer; the corresponding transference plan π can be written as

$$\pi = \frac{1}{n} \sum_{ij} a_{ij}\, \delta_{(x_i, y_j)},$$

and its support S is the set of all couples (x_i, y_j) such that $a_{ij} > 0$.

Assume that S is not cyclically monotone: Then there exists $(x_{i_1}, y_{j_1}), \ldots, (x_{i_N}, y_{j_N})$ in S such that

$$c(x_{i_1}, y_{j_2}) + c(x_{i_2}, y_{j_3}) + \ldots + c(x_{i_N}, y_{j_1}) < c(x_{i_1}, y_{j_1}) + \ldots + c(x_{i_N}, y_{j_N}). \tag{5.17}$$

Let $a := \min(a_{i_1, j_1}, \ldots, a_{i_N, j_N}) > 0$. Define a new transference plan $\widetilde{\pi}$ by the formula

$$\widetilde{\pi} = \pi + \frac{a}{n} \sum_{\ell=1}^{N} \big(\delta_{(x_{i_\ell}, y_{j_{\ell+1}})} - \delta_{(x_{i_\ell}, y_{j_\ell})} \big).$$

It is easy to check that this has the correct marginals, and by (5.17) the cost associated with $\widetilde{\pi}$ is strictly less than the cost associated with π. This is a contradiction, so S is indeed c-cyclically monotone!

Step 2: *If c is continuous, then there is a cyclically monotone transference plan.*

To prove this, consider sequences of independent random variables $x_i \in \mathcal{X}$, $y_j \in \mathcal{Y}$, with respective law μ, ν. According to the law of large numbers for empirical measures (sometimes called fundamental theorem of statistics, or Varadarajan's theorem), one has, with probability 1,

$$\mu_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} \longrightarrow \mu, \qquad \nu_n := \frac{1}{n} \sum_{j=1}^{n} \delta_{y_j} \longrightarrow \nu \tag{5.18}$$

as $n \to \infty$, in the sense of weak convergence of measures. In particular, by Prokhorov's theorem, (μ_n) and (ν_n) are tight sequences.

For each n, let π_n be a cyclically monotone transference plan between μ_n and ν_n. Let $\varepsilon > 0$ be arbitrary; by tightness there are compact sets $K_\varepsilon \subset \mathcal{X}$ and $L_\varepsilon \subset \mathcal{Y}$ such that for all n, $\mu_n[\mathcal{X} \setminus K_\varepsilon] \le \varepsilon$, $\nu_n[\mathcal{Y} \setminus L_\varepsilon] \le \varepsilon$; then

$$\begin{aligned}\pi_n\big[(\mathcal{X} \times \mathcal{Y}) \setminus (K_\varepsilon \times L_\varepsilon)\big] &\le \pi_n\big[\mathcal{X} \times (\mathcal{Y} \setminus L_\varepsilon)\big] + \pi_n\big[(\mathcal{X} \setminus K_\varepsilon) \times \mathcal{Y}\big] \\ &= \nu_n[\mathcal{Y} \setminus L_\varepsilon] + \mu_n[\mathcal{X} \setminus K_\varepsilon] \le 2\varepsilon.\end{aligned}$$

Since $K_\varepsilon \times L_\varepsilon$ is compact, this proves the tightness of the sequence (π_n). By Prokhorov's theorem again, there is a subsequence, still denoted (π_n), which converges weakly to some probability measure π, i.e.

$$\int h(x,y)\, d\pi_n(x,y) \longrightarrow \int h(x,y)\, d\pi(x,y)$$

for all bounded continuous functions h on $\mathcal{X} \times \mathcal{Y}$. By applying the previous identity with $h(x,y) = f(x)$ and $h(x,y) = g(y)$, we see that π has marginals μ and ν, so this is an admissible transference plan between μ and ν.

For each n, the cyclical monotonicity of π_n implies that for all N and $\pi_n^{\otimes N}$-almost all $(x_1,y_1),\dots,(x_N,y_N)$, the inequality (5.1) is satisfied; in other words, $\pi_n^{\otimes N}$ is concentrated on the set $\mathcal{C}(N)$ of all $((x_1,y_1),\dots,(x_N,y_N)) \in (\mathcal{X} \times \mathcal{Y})^N$ satisfying (5.1). Since c is continuous, $\mathcal{C}(N)$ is a *closed* set, so the weak limit $\pi^{\otimes N}$ of $\pi_n^{\otimes N}$ is also concentrated on $\mathcal{C}(N)$. Let $\Gamma = \mathrm{Spt}\,\pi$ (Spt stands for "support"), then

$$\Gamma^N = (\mathrm{Spt}\,\pi)^N = \mathrm{Spt}(\pi^{\otimes N}) \subset \mathcal{C}(N),$$

and since this holds true for all N, Γ is c-cyclically monotone.

Step 3: *If c is continuous and π is c-cyclically monotone, there is a c-convex ψ such that $\partial_c\psi$ contains the support of π.*

Let again Γ denote the support of π (this is a closed set). Pick any $(x_0,y_0) \in \Gamma$, and define

$$\begin{aligned}\psi(x) := \sup_{m\in\mathbb{N}} \sup\Big\{ &\big[c(x_0,y_0) - c(x_1,y_0)\big] + \big[c(x_1,y_1) - c(x_2,y_1)\big] \\ &+ \cdots + \big[c(x_m,y_m) - c(x,y_m)\big];\quad (x_1,y_1),\dots,(x_m,y_m) \in \Gamma \Big\}. \quad (5.19)\end{aligned}$$

By applying the definition with $m = 1$ and $(x_1,y_1) = (x_0,y_0)$, one immediately sees that $\psi(x_0) \ge 0$. On the other hand, $\psi(x_0)$ is the supremum of all the quantities $[c(x_0,y_0) - c(x_1,y_0)] + \dots + [c(x_m,y_m) - c(x_0,y_m)]$ which by cyclical monotonicity are all nonpositive. So actually $\psi(x_0) = 0$. In fact this is the only place in this Step where c-cyclically monotonicity will be used!

By renaming y_m as y, obviously

$$\begin{aligned}\psi(x) = \sup_{y\in\mathcal{Y}} \sup_{m\in\mathbb{N}} \sup_{(x_1,y_1),\dots,(x_{m-1},y_{m-1}),x_m} \Big\{ &\big[c(x_0,y_0) - c(x_1,y_0)\big] + \big[c(x_1,y_1) - c(x_2,y_1)\big] \\ &+ \cdots + \big[c(x_m,y) - c(x,y)\big];\quad (x_1,y_1),\dots,(x_m,y) \in \Gamma \Big\}. \quad (5.20)\end{aligned}$$

So $\psi(x) = \sup_y[\zeta(y) - c(x,y)]$, if ζ is defined by

$$\begin{aligned}\zeta(y) = \sup \Big\{ &\big[c(x_0,y_0) - c(x_1,y_0)\big] + \big[c(x_1,y_1) - c(x_2,y_1)\big] + \cdots + c(x_m,y); \\ &m \in \mathbb{N}, (x_1,y_1),\dots,(x_m,y) \in \Gamma \Big\} \quad (5.21)\end{aligned}$$

(with the convention that $\zeta = -\infty$ out of $\mathrm{proj}_{\mathcal{Y}}(\Gamma)$). Thus ψ is a c-convex function.

Now let $(\overline{x}, \overline{y}) \in \Gamma$. By choosing $\overline{x} = x_m$, $\overline{y} = y_m$ in the definition of ψ,

$$\psi(x) \geq \sup_m \Big\{ \Big(\sup_{(x_1,y_1),\dots,(x_{m-1},y_{m-1})} [c(x_0,y_0) - c(x_1,y_0)] + \cdots + [c(x_{m-1},y_{m-1}) - c(\overline{x},y_{m-1})] \Big) + [c(\overline{x},\overline{y}) - c(x,\overline{y})], \Big\}.$$

In the definition of ψ, it does not matter whether one takes the supremum over $m-1$ or over m variables, since one also takes the supremum over m. So the previous inequality can be recast as

$$\psi(x) \geq \psi(\overline{x}) + c(\overline{x},\overline{y}) - c(x,\overline{y}).$$

In particular, $\psi(x) + c(x,\overline{y}) \geq \psi(\overline{x}) + c(\overline{x},\overline{y})$. Taking the infimum over $x \in \mathcal{X}$ in the left-hand side, we deduce that

$$\psi^c(\overline{y}) \geq \psi(\overline{x}) + c(\overline{x},\overline{y}).$$

Since the reverse inequality is always satisfied, actually

$$\psi^c(\overline{y}) = \psi(\overline{x}) + c(\overline{x},\overline{y}),$$

and this means precisely that $(\overline{x},\overline{y}) \in \partial_c\psi$. So Γ does lie in the c-subdifferential of ψ.

Step 4: *If c is continuous and bounded, then there is duality.*

Let $\|c\| := \sup c(x,y)$. By Steps 2 and 3, there exists a transference plan π whose support is included in $\partial_c\psi$ for some c-convex ψ, which was constructed "explicitly" in Step 3. Let $\phi = \psi^c$.

From (5.19), $\psi = \sup \psi_m$, where each ψ_m is a supremum of continuous functions, and therefore lower semi-continuous. In particular, ψ is measurable. (Note: A lower semi-continuous function on a Polish space is always measurable, even if it is obtained as a supremum of uncountably many continuous functions; in fact it can always be written as a supremum of countably many continuous functions!) The same can be said of ϕ.

Next we check that ψ, ϕ are bounded. Let $(x_0,y_0) \in \partial_c\psi$ be such that $\psi(x_0) < +\infty$; then necessarily $\phi(y_0) > -\infty$. So, for any $x \in \mathcal{X}$,

$$\psi(x) = \sup_y \, [\phi(y) - c(x,y)] \;\geq\; \phi(y_0) - c(x,y_0) \;\geq\; \phi(y_0) - \|c\|;$$

$$\phi(y) = \inf_x \, [\psi(x) + c(x,y)] \;\leq\; \psi(x_0) + c(x_0,y) \;\leq\; \psi(x_0) + \|c\|.$$

Re-injecting these bounds in the identities $\psi = \phi^c$, $\phi = \psi^c$, we get

$$\psi(x) \;\leq\; \sup_y \, \phi(y) \leq \;\psi(x_0) + \|c\|;$$

$$\phi(y) \;\geq\; \inf_x \, \psi(x) \geq \;\phi(y_0) - \|c\|.$$

So both ψ and ϕ are bounded from above and below.

Thus we can integrate ϕ, ψ against μ, ν respectively, and, by the marginal condition,

$$\int \phi(y)\, d\nu(y) - \int \psi(x)\, d\mu(x) = \int \big[\phi(y) - \psi(x)\big]\, d\pi(x,y).$$

Since $\phi(y) - \psi(x) = c(x,y)$ on the support of π, the latter quantity equals $\int c(x,y)\, d\pi(x,y)$. It follows that (5.4) is actually an equality, which proves the duality.

Step 5: *If c is lower semicontinuous, then there is duality.*

Since c is lower semi-continuous, we can write

$$c(x,y) = \lim_{k\to\infty} c_k(x,y),$$

where $(c_k)_{k\in\mathbb{N}}$ is a nondecreasing sequence of bounded, uniformly continuous functions. To see this, just choose

$$c_k(x,y) = \inf_{(x',y')} \Big\{ \min\Big(c(x',y'), k\Big) + k\Big[d(x,x') + d(y,y')\Big] \Big\};$$

note that c_k is k-Lipschitz, nondecreasing in k, and satisfies $0 \le c_k(x,y) \le \min(c(x,y),k)$. (It is instructive to understand exactly where the lower semi-continuity assumption is used in this statement.)

By Step 4, for each k we can find π_k, ϕ_k, ψ_k such that ψ_k is bounded and c-convex, $\phi_k = (\psi_k)^c$, and

$$\int c_k(x,y)\, d\pi_k(x,y) = \int \phi_k(y)\, d\nu(y) - \int \psi_k(x)\, d\mu(x).$$

Since c_k is no larger than c, the constraint $\phi_k(y) - \psi_k(x) \le c_k(x,y)$ implies $\phi_k(y) - \psi_k(x) \le c(x,y)$; so all (ϕ_k, ψ_k) are admissible in the dual problem with cost c. Moreover, for each k the functions ϕ_k and ψ_k are uniformly continuous because c itself is uniformly continuous.

By Lemma 4.3, $\Pi(\mu,\nu)$ is weakly sequentially compact. Thus, up to extraction of a subsequence, we can assume that π_k converges to some $\widetilde{\pi} \in \Pi(\mu,\nu)$. For all indices $\ell \le k$, we have $c_\ell \le c_k$, so

$$\begin{aligned}
\int c_\ell\, d\widetilde{\pi} &= \lim_{k\to\infty} \int c_\ell\, d\pi_k \\
&\le \limsup_{k\to\infty} \int c_k\, d\pi_k \\
&= \limsup_{k\to\infty} \left(\int \phi_k(y)\, d\nu(y) - \int \psi_k(x)\, d\mu(x) \right).
\end{aligned}$$

On the other hand, by monotone convergence,

$$\int c\, d\widetilde{\pi} = \lim_{\ell\to\infty} \int c_\ell\, d\widetilde{\pi}.$$

So

$$\begin{aligned}
\inf_{\Pi(\mu,\nu)} \int c\, d\pi \le \int c\, d\widetilde{\pi} &\le \limsup_{k\to\infty} \left(\int \phi_k(y)\, d\nu(y) - \int \psi_k(x)\, d\mu(x) \right) \\
&\le \inf_{\Pi(\mu,\nu)} \int c\, d\pi.
\end{aligned}$$

Moreover,

$$\int \phi_k(y)\, d\nu(y) - \int \psi_k(x)\, d\mu(x) \longrightarrow \inf_{\Pi(\mu,\nu)} \int c\, d\pi. \tag{5.22}$$

Since each pair (ψ_k, ϕ_k) lies in $C_b(\mathcal{X}) \times C_b(\mathcal{Y})$, the duality also holds with bounded continuous (and even Lipschitz) test functions, as claimed in Theorem 5.9(i). □

Proof of Theorem 5.9, Part (ii). From now on, I shall assume that the optimal transport cost $C(\mu, \nu)$ is finite, and that c is real-valued. Part (ii) will be established in the following way: (a) $\Rightarrow$ (b) $\Rightarrow$ (c) $\Rightarrow$ (d) $\Rightarrow$ (a) $\Rightarrow$ (e) $\Rightarrow$ (b). There seems to be some redundancy in this chain of implications, but this is because the implications (a) $\Rightarrow$ (c) will be used to construct the set Γ appearing in (e).

(a) $\Rightarrow$ (b): Let π be an optimal plan, and let $(\phi_k, \psi_k)_{k \in \mathbb{N}}$ be as in Step 5 of the proof of Part (i). Since the optimal transport cost is finite by assumption, the cost function c belongs to $L^1(\pi)$. From (5.22) and the marginal property of π,

$$\int \Big[c(x,y) - \phi_k(y) + \psi_k(x) \Big] \, d\pi(x,y) \longrightarrow 0$$

as $k \to \infty$, which means that $c(x,y) - \phi_k(y) + \psi_k(x)$ converges to 0 in $L^1(\mathcal{X} \times \mathcal{Y}, \pi)$ as $k \to \infty$. Up to choosing a subsequence, we can assume that the convergence is almost sure; then $\phi_k(y_i) - \psi_k(x_i)$ converges to $c(x_i, y_i)$, $\pi(dx_i \, dy_i)$-almost surely, as $k \to \infty$. By passing to the limit in the inequality

$$\sum_{i=1}^N c(x_i, y_{i+1}) \geq \sum_{i=1}^N [\phi_k(y_{i+1}) - \psi_k(x_i)] = \sum_{i=1}^N [\phi_k(y_i) - \psi_k(x_i)]$$

(where by convention $y_{N+1} = y_1$) we see that, $\pi^{\otimes N}$-almost surely,

$$\sum_{i=1}^N c(x_i, y_{i+1}) \geq \sum_{i=1}^N c(x_i, y_i). \tag{5.23}$$

The conclusion so far is that there is a set $\Gamma_N \subset (\mathcal{X} \times \mathcal{Y})^N$ of N-tuples $((x_1, y_1), \ldots, (x_N, y_N))$, of full measure for $\pi^{\otimes N}$, satisfying (5.23). Let $\mathrm{proj}_k((x_i, y_i)_{1 \leq i \leq N}) := (x_k, y_k)$ be the projection on the factor k of $(\mathcal{X} \times \mathcal{Y})^N$. It is easy to check that $\Gamma := \cap_{1 \leq k \leq N} \mathrm{proj}_k(\Gamma_N)$ is a c-cyclically monotone set which has full π-measure; so π is indeed c-cyclically monotone.

(b) $\Rightarrow$ (c): Let π be a cyclically monotone transference plan. The function ψ can be constructed just as in Step 3 of the proof of Part (i), only with some differences. First, Γ is not necessarily closed; it is just a Borel set such that $\pi[\Gamma] = 1$. (If Γ is not Borel, make it Borel by modifying it on a negligible set.) With this in mind, define, as in Step 3 of Part (i),

$$\begin{aligned} \psi(x) := \sup_{m \in \mathbb{N}} \sup \Big\{ & \big[c(x_0, y_0) - c(x_1, y_0) \big] + \big[c(x_1, y_1) - c(x_2, y_1) \big] \\ & + \cdots + \big[c(x_m, y_m) - c(x, y_m) \big]; \quad (x_1, y_1), \ldots, (x_m, y_m) \in \Gamma \Big\}. \end{aligned} \tag{5.24}$$

From its definition, for any $x \in \mathcal{X}$,

$$\psi(x) \geq c(x_0, y_0) - c(x, y_0) > -\infty.$$

(Here the assumption of c being real-valued is useful.) Then there is no difficulty in proving, as in Step 3, that $\psi(x_0) = 0$, that ψ is c-convex, and that π is concentrated on $\partial_c \psi$.

The rest of this step will be devoted to the *measurability* of ψ, ψ^c and $\partial_c \psi$. These are surprisingly subtle issues, which do not arise if c is continuous; so the reader which only cares for a continuous cost function might go directly to the next step.

First, the measurability of ψ is not clear at all from formula (5.24): This is typically an uncountable supremum of *upper* semi-continuous functions, and there is no a priori reason for this to be Borel measurable.

Since c is lower semi-continuous and nonnegative, there is an increasing sequence $(c_\ell)_{\ell\in\mathbb{N}}$ of continuous nonnegative functions, such that $c_\ell(x,y)$ converges to $c(x,y)$ as $\ell\to\infty$, for all (x,y). By Egorov's theorem, for each $k\in\mathbb{N}$ there is a Borel set E_k with $\pi[E_k]\le 1/k$, such that the convergence of c_ℓ to c is uniform on $\Gamma\setminus E_k$. Since π (as any probability measure on a Polish space) is regular, we can find a *compact* set $\Gamma_k\subset\Gamma\setminus E_k$, such that $\pi[\Gamma_k]\ge 1-2/k$. There is no loss of generality in assuming that the sets Γ_k are increasing in k.

On each Γ_k, the sequence (c_ℓ) converges uniformly and monotonically to c; in particular c is continuous on Γ_k. Furthermore, since π is obviously concentrated on the union of all Γ_k, there is no loss of generality in assuming that $S=\cup\Gamma_k$. We may also assume that $(x_0,y_0)\in\Gamma_1$.

Now, let x be given in $\mathcal{X}$, and for each k,ℓ,m, let

$$F_{m,k,\ell}\big(x_0,y_0,\dots,x_m,y_m\big) := \big[c(x_0,y_0)-c_\ell(x_1,y_0)\big]+\big[c(x_1,y_1)-c_\ell(x_2,y_1)\big] \\ +\ \cdots+\ \big[c(x_m,y_m)-c_\ell(x,y_m)\big],$$

for $(x_0,y_0,\dots,x_m,y_m)\in\Gamma_k^m$. It is clear that $F_{k,\ell,m}$ is a continuous function (because c_ℓ is continuous on $\mathcal{X}\times\mathcal{X}$, and c is continuous on Γ_k). It is defined on the compact set Γ_k^m, and it is nonincreasing as a function of ℓ, with

$$\lim_{\ell\to\infty} F_{m,k,\ell}=F_{m,k},$$

where

$$F_{m,k}\big(x_0,y_0,\dots,x_m,y_m\big) := \big[c(x_0,y_0)-c(x_1,y_0)\big]+\big[c(x_1,y_1)-c(x_2,y_1)\big] \\ +\ \cdots+\ \big[c(x_m,y_m)-c(x,y_m)\big].$$

Now I claim that

$$\lim_{\ell\to\infty}\ \sup_{\Gamma_k^m}\ F_{m,k,\ell}\ =\ \sup_{\Gamma_k^m}\ F_{m,k}. \tag{5.25}$$

Indeed, by compactness, for each $\ell\in\mathbb{N}$ there is $X_\ell\in\Gamma_k^m$ such that

$$\sup_{\Gamma_k^m} F_{m,k,\ell}=F_{m,k,\ell}(X_\ell);$$

and up to extraction of a subsequence, one may assume that X_ℓ converges to some X. Then by monotonicity, for any $\ell'\le\ell$,

$$\sup_{\Gamma_k^m}\ F_{m,k,\ell}=F_{m,k,\ell}(X_\ell)\ \le\ F_{m,k,\ell'}(X_\ell);$$

and if one lets $\ell\to\infty$, with ℓ' fixed, one obtains

$$\limsup_{\ell\to\infty}\ \sup_{\Gamma_k^m}\ F_{m,k,\ell}\le\ F_{m,k,\ell'}(X).$$

Now let $\ell'\to\infty$, to get

$$\limsup_{\ell\to\infty}\ \sup_{\Gamma_k^m}\ F_{m,k,\ell}\ \le\ F_{m,k}(X)\ \le\ \sup_{\Gamma_k^m}\ F_{m,k}.$$

The converse inequality

$$\sup_{\Gamma_k^m} F_{m,k} \;\le\; \liminf_{\ell\to\infty} \sup_{\Gamma_k^m} F_{m,k,\ell}$$

is obvious because $F_{m,k} \le F_{m,k,\ell}$; so (5.25) is proven.

To summarize: If we let

$$\psi_{m,k,\ell}(x) := \sup \Big\{ \big[c(x_0,y_0) - c_\ell(x_1,y_0)\big] + \big[c(x_1,y_1) - c_\ell(x_2,y_1)\big] \\ + \cdots + \big[c(x_m,y_m) - c_\ell(x,y_m)\big]; \quad (x_1,y_1),\dots,(x_m,y_m) \in \Gamma_k \Big\},$$

then we have

$$\lim_{\ell\to\infty} \psi_{m,k,\ell}(x) \;=\; \sup \Big\{ \big[c(x_0,y_0) - c_\ell(x_1,y_0)\big] + \big[c(x_1,y_1) - c_\ell(x_2,y_1)\big] \\ + \cdots + \big[c(x_m,y_m) - c_\ell(x,y_m)\big]; \quad (x_1,y_1),\dots,(x_m,y_m) \in \Gamma_k \Big\}.$$

It follows easily that, for each x,

$$\psi(x) \;= \sup_{m\in\mathbb{N}} \sup_{k\in\mathbb{N}} \lim_{\ell\to\infty} \psi_{m,k,\ell}(x).$$

Since $\psi_{m,k,\ell}(x)$ is lower semi-continuous in x (as a supremum of continuous functions of x), it follows that ψ itself is measurable.

The measurability of $\phi := \psi^c$ is subtle also, and at the present level of generality it is not clear that this function is really Borel measurable. However, it can be modified on a ν-negligible set so as to become measurable. Indeed, $\phi(y) - \psi(x) = c(x,y)$ π-almost surely, so if one disintegrates $\pi(dx\,dy)$ as $\pi(dx|y)\,\nu(dy)$, then $\phi(y)$ coincides, $\nu(dy)$-almost surely, with the Borel function $\widetilde{\phi}(y) := \int_{\mathcal{X}} [\psi(x) + c(x,y)]\,\pi(dx|y)$.

Let then Z be a Borel set of zero ν-measure such that $\widetilde{\phi} = \phi$ outside of Z. Then the subdifferential $\partial_c\psi$ coincides, out of the π-negligible set $\mathcal{X}\times Z$, with the measurable set $\{(x,y) \in \mathcal{X}\times\mathcal{Y};\ \widetilde{\phi}(y) - \psi(x) = c(x,y)\}$. The conclusion is that $\partial_c\psi$ can be modified on a π-negligible set so as to be Borel measurable.

(c) ⇒ (d): Just let $\phi = \psi^c$.

(d) ⇒ (a): Let (ψ,ϕ) be a pair of admissible functions, and let π be a transference plan such that $\phi - \psi = c$ π-almost surely. The goal is to show that π is optimal. The main difficulty lies in the fact that ψ and ϕ need not be *separately* integrable. This problem will be circumvented by a careful truncation procedure. For $n \in \mathbb{N}$, $w \in \mathbb{R}\cup\{\pm\infty\}$, define

$$T_n(w) = \begin{cases} w & \text{if } |w| \le n \\ n & \text{if } w > n \\ -n & \text{if } w < -n, \end{cases}$$

and

$$\xi(x,y) := \phi(y) - \psi(x); \qquad \xi_n(x,y) := T_n(\phi(y)) - T_n(\psi(x)).$$

In particular, $\xi_0 = 0$. It is easily checked that ξ_n converges *monotonically* to ξ; more precisely,

- $\xi_n(x,y)$ remains equal to 0 if $\xi(x,y) = 0$;
- $\xi_n(x,y)$ increases to $\xi(x,y)$ if the latter quantity is positive;
- $\xi_n(x,y)$ decreases to $\xi(x,y)$ if the latter quantity is negative.

As a consequence, $\xi_n \le (\xi_n)_+ \le \xi_+ \le c$. So $(T_n\phi, T_n\psi)$ is an admissible pair in the dual Kantorovich problem, and

$$\int \xi_n \, d\pi = \int (T_n\phi)\, d\nu - \int (T_n\psi)\, d\mu \le \sup_{\phi'-\psi'\le c} \left(\int \phi' \, d\mu - \int \psi' \, d\nu \right). \tag{5.26}$$

On the other hand, by monotone convergence and since ξ coincides with c outside of a π-negligible set,

$$\int_{\xi\ge 0} \xi_n \, d\pi \xrightarrow[n\to\infty]{} \int_{\xi\ge 0} \xi \, d\pi = \int c \, d\pi;$$

$$\int_{\xi<0} \xi_n \, d\pi \xrightarrow[n\to\infty]{} \int_{\xi<0} \xi \, d\pi = 0.$$

This and (5.26) imply that

$$\int c \, d\pi \le \sup_{\phi'-\psi'\le c} \left(\int \phi' \, d\mu - \int \psi' \, d\nu \right);$$

so π is optimal.

Now we can construct the set Γ appearing in (ii). By Theorem 4.1, there is at least one optimal transference plan, say $\widetilde{\pi}$. From the implication (a) $\Rightarrow$ (c), there is some $\widetilde{\psi}$ such that $\widetilde{\pi}$ is concentrated on $\partial_c\widetilde{\psi}$; just choose $\Gamma := \partial_c\widetilde{\psi}$.

(a) $\Rightarrow$ (e): Let $\widetilde{\pi}$ be the optimal plan used to construct Γ, and let π be another optimal plan. Since π and $\widetilde{\pi}$ have the same cost and same marginals,

$$\begin{aligned} \int c \, d\pi = \int c \, d\widetilde{\pi} &= \lim_{n\to\infty} \int (T_n\phi - T_n\psi)\, d\widetilde{\pi} \\ &= \lim_{n\to\infty} \int (T_n\phi - T_n\psi)\, d\pi, \end{aligned}$$

where T_n is the same truncation operator that was already used in the proof of (d) $\Rightarrow$ (a). So

$$\int [c(x,y) - T_n\phi + T_n\psi] \, d\pi \xrightarrow[n\to\infty]{} 0. \tag{5.27}$$

As before, define $\xi(x,y) := \phi(y) - \psi(x)$; then by monotone convergence,

$$\int_{\xi\ge 0} [c - T_n\phi + T_n\psi] \, d\pi \xrightarrow[n\to\infty]{} \int_{\xi\ge 0} (c-\xi) \, d\pi;$$

$$\int_{\xi<0} [c - T_n\phi + T_n\psi] \, d\pi \xrightarrow[n\to\infty]{} \int_{\xi<0} (c-\xi) \, d\pi.$$

Since $\xi \le c$, the integrands here are nonnegative and both integrals make sense in $\mathbb{R} \cup \{+\infty\}$. So by adding the two limits and using (5.27) we get

$$\int (c-\xi) \, d\pi = \lim_{n\to\infty} \int [c - T_n\phi + T_n\psi] = 0.$$

Since $\xi \le c$, this proves that c coincides π-almost surely with ξ, which was the desired conclusion.

(e) $\Rightarrow$ (b): This is obvious since Γ is cyclically monotone by assumption. □

Proof of Theorem 5.9, Part (iii). Let π be optimal, and let ψ be a c-convex function such that π is concentrated on $\partial_c\psi$. Let $\phi := \psi^c$. The goal is to show that under the assumption $c \le c_{\mathcal{X}} + c_{\mathcal{Y}}$, (ψ, ϕ) solves the dual Kantorovich problem.

The point is to show that ψ and ϕ are integrable. To show this we repeat the estimates of Step 4 in the proof of Part (i), with some variants: After securing (x_0, y_0) such that $\phi(y_0)$, $\psi(x_0)$, $c_{\mathcal{X}}(x_0)$ and $c_{\mathcal{Y}}(y_0)$ are finite, we write

$$\psi(x) + c_{\mathcal{X}}(x) = \sup_y \big[\phi(y) - c(x,y) + c_{\mathcal{X}}(x)\big] \geq \sup_y [\phi(y) - c_{\mathcal{Y}}(y)] \geq \phi(y_0) - c_{\mathcal{Y}}(y_0);$$

$$\phi(y) - c_{\mathcal{Y}}(y) = \inf_x \big[\psi(x) + c(x,y) - c_{\mathcal{Y}}(y)\big] \leq \inf_x [\psi(x) + c_{\mathcal{X}}(x)] \leq \psi(x_0) + c_{\mathcal{X}}(x_0).$$

So ψ is bounded below by the μ-integrable function $\phi(y_0) - c_{\mathcal{Y}}(y_0) - c_{\mathcal{X}}(x)$, and ϕ is bounded above by the ν-integrable function $\psi(x_0) + c_{\mathcal{X}}(x_0) + c_{\mathcal{Y}}(y)$. So both $-\int \psi \, d\mu$ and $\int \phi \, d\nu$ make sense in $\mathbb{R} \cup \{+\infty\}$. Since their sum is $\int (\phi - \psi) \, d\pi = \int c \, d\pi < +\infty$, both integrals are finite. So

$$\int c \, d\pi = \int \phi \, d\nu - \int \psi \, d\mu,$$

and it follows from Part (i) of the theorem that both π and (ψ, ϕ) are optimal, respectively in the original and the dual Kantorovich problems.

To prove the last part of (iii), assume that c is continuous; then the subdifferential of any c-convex function is a closed c-cyclically monotone set.

Let π be an arbitrary optimal transference plan, and (ψ, ϕ) an optimal pair of prices. We know that (ψ, ψ^c) is optimal in the dual Kantorovich problem, so

$$\int c(x,y) \, d\pi(x,y) = \int \psi^c \, d\nu - \int \psi \, d\mu.$$

Thanks to the marginal condition, this be rewritten as

$$\int \big[\psi^c(y) - \psi(x) - c(x,y)\big] \, d\pi(x,y) = 0.$$

Since the integrand is nonnegative, it follows that π is concentrated on the set of pairs (x,y) such that $\psi^c(y) - \psi(x) - c(x,y) = 0$, that is precisely the subdifferential of ψ. Thus any optimal transference plan is concentrated on the subdifferential of any optimal ψ. So if Γ is defined as the intersection of all subdifferentials of optimal functions ψ, then Γ also contains the supports of all optimal plans.

Conversely, if $\widetilde{\pi} \in \Pi(\mu, \nu)$ is a transference plan concentrated on Γ, then $\int c \, d\widetilde{\pi} = \int [\psi^c - \psi] \, d\widetilde{\pi} = \int \psi^c \, d\nu - \int \psi \, d\mu$, so $\widetilde{\pi}$ is optimal. Similarly, if $\widetilde{\psi}$ is a c-convex function such that $\partial_c \widetilde{\psi}$ contains Γ, then by the previous estimates $\widetilde{\psi}$ and $\widetilde{\psi}^c$ are integrable against μ and ν respectively, and $\int c \, d\pi = \int [\widetilde{\psi}^c - \widetilde{\psi}] \, d\pi = \int \widetilde{\psi}^c \, d\nu - \int \widetilde{\psi} \, d\mu$, so $(\widetilde{\psi}, \widetilde{\psi}^c)$ is optimal. This concludes the proof. □

Restriction property

The dual side of the Kantorovich problem also behaves well with respect to restriction, as shown by the next results (who may look natural, if not obvious, to the reader).

Theorem 5.16 (Restriction theorem for the price function). *Let $\mathcal{X}, \mathcal{Y}$ be two Polish spaces, and $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a lower semi-continuous cost function, bounded below, and let $\psi : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ be a c-convex function. Let $\mathcal{X}' \subset \mathcal{X}$ and $\mathcal{Y}' \subset \mathcal{Y}$ be closed sets, and let c' be the restriction of c to $\mathcal{X}' \times \mathcal{Y}'$. Let further Γ' be a measurable c'-cyclically monotone subset included in $\partial_c \psi$. Then, there is a c'-convex function $\psi' : \mathcal{X}' \to \mathbb{R} \cup \{+\infty\}$ such that ψ' coincides with ψ on* $\mathrm{proj}_{\mathcal{X}}(\Gamma')$.

Corollary 5.17 (Restriction for the duality theorem). *Let $\mathcal{X}, \mathcal{Y}$ be two Polish spaces, and $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a lower semi-continuous cost function, bounded below, let $\mu \in P(\mathcal{X})$ and $\nu \in P(\mathcal{Y})$ be probability measures such that the optimal total cost $C(\mu, \nu)$ is finite. Let π be an optimal transference plan, and let ψ be such that π is supported on $\partial_c \psi$. Let $\widetilde{\pi}$ be a measure satisfying $\widetilde{\pi} \le \pi$, and $\pi' := \widetilde{\pi}/\widetilde{\pi}[\mathcal{X} \times \mathcal{Y}]$. Let further μ' and ν' be the marginals of π', and let $\mathcal{X}' := \mathrm{Spt}\, \mu'$, $\mathcal{Y}' := \mathrm{Spt}\, \nu'$. Let c' be the restriction of the cost function c to $\mathcal{X}' \times \mathcal{Y}'$. Then there is a c'-convex function ψ' on $\mathcal{X}'$ such that π' is concentrated on $\partial_{c'} \psi'$ and ψ' coincides with ψ μ'-almost surely.*

Proof of Theorem 5.16. Let $\phi := \psi^c$, and for any $x \in \mathcal{X}'$ define

$$\psi'(x) := \sup_{y \in \mathcal{Y}'} \big[\widetilde{\phi}(y) - c(x,y)\big],$$

where

$$\widetilde{\phi}(y) := \begin{cases} \phi(y) & \text{if there is } x' \in \mathcal{X}' \text{ such that } (\widetilde{x}, y) \in \Gamma'; \\ -\infty & \text{otherwise.} \end{cases}$$

By definition, ψ' is a c'-convex function, and obviously

$$\forall x \in \mathcal{X}', \qquad \psi'(x) \le \sup_{y \in \mathcal{Y}} \big[\phi(y) - c(x,y)\big] = \psi(x).$$

Then, for any $x \in \mathrm{proj}_{\mathcal{X}}(\Gamma')$, there is y such that $(x,y) \in \Gamma' \subset \partial_c \psi$, and so $\psi'(x) \ge \phi(y) = \psi(x)$. So ψ' coincides on ψ on $\mathrm{proj}_{\mathcal{X}}(\Gamma')$.

The conclusion follows since $\mu' = (\mathrm{proj}_{\mathcal{X}})_{\#} \pi'$ is concentrated on $\mathrm{proj}_{\mathcal{X}}(\Gamma')$. Note that $\mathrm{proj}_{\mathcal{X}}(\Gamma')$ is not a priori measurable, yet by regularity we can find an increasing sequence of compact sets $(K_\ell)_{\ell \in \mathbb{N}}$ included in Γ' such that π' is concentrated on $\cup K_\ell$; and then μ' is concentrated on $\cup \mathrm{proj}_{\mathcal{X}}(K_\ell)$, which is measurable as a countable union of compact sets, and still included in $\mathrm{proj}_{\mathcal{X}}(\Gamma')$. □

Proof of Corollary 5.17. It follows from Theorem 5.9 that π is concentrated on $\partial_c \psi$; so π' is concentrated on $\Gamma' := \partial_c \psi \cap (\mathcal{X}' \times \mathcal{Y}')$. Then the conclusion follows from Theorem 5.16. □

Application: Stability

An important consequence of Theorem 5.9 is the stability of optimal transference plans.

Theorem 5.18 (Stability of optimal transport). *Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces, and let $c(x,y)$ be a real-valued continuous cost function,* $\inf c > -\infty$. *Let c_k be a family of continuous cost functions converging uniformly to c on $\mathcal{X} \times \mathcal{Y}$. Let $(\mu_k)_{k \in \mathbb{N}}$ and $(\nu_k)_{k \in \mathbb{N}}$ be sequences of probability measures on $\mathcal{X}$ and $\mathcal{Y}$ respectively. Assume that μ_k converges to μ (resp. ν_k converges to ν) weakly. For each k, let further π_k be an optimal transport plan between μ_k and ν_k. If*

$$\forall k \in \mathbb{N}, \qquad \int c_k(x,y)\, d\pi_k < +\infty,$$

then, up to extraction of a subsequence, π_k converges weakly to a c-monotone transference plan $\pi \in \Pi(\mu, \nu)$.

If moreover

$$\liminf_{k \in \mathbb{N}} \int c_k\, d\pi_k < +\infty,$$

then the optimal total transport cost $C(\mu, \nu)$ between μ and ν is finite, and π is an optimal transference plan.

Corollary 5.19 (Compactness of the set of optimal plans). *Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces, and let $c(x, y)$ be a real-valued continuous cost function,* $\inf c > -\infty$. *Let $\mathcal{K}$ and $\mathcal{L}$ be two compact subsets of $P(\mathcal{X})$ and $P(\mathcal{Y})$ respectively. Then the set of optimal transference plans π between $\mu \in \mathcal{K}$ and $\nu \in \mathcal{L}$ is itself compact in $P(\mathcal{X} \times \mathcal{Y})$.*

Proof of Theorem 5.18. Since μ_k and ν_k are convergent sequences, by Prokhorov's theorem they form tight sets, and then by Lemma 4.3 the measures π_k all lie in a tight set of $\mathcal{X} \times \mathcal{Y}$; so we can extract a further subsequence, still denoted (π_k) for simplicity, which converges weakly to some $\pi \in \Pi(\mu, \nu)$.

To prove that π is c-monotone, the argument is essentially the same as in Step 2 of the proof of Theorem 5.9(i). Indeed, by Theorem 5.9, each π_k is concentrated on a c_k-cyclically monotone set; so $\pi_k^{\otimes N}$ is concentrated on the set $\mathcal{C}_k(N)$ of $((x_1, y_1), \dots, (x_N, y_N))$ such that

$$\sum_{1 \le k \le N} c_k(x_i, y_i) \le \sum_{1 \le k \le N} c_k(x_i, y_{i+1}),$$

where as usual $y_{N+1} = y_1$. So if $\varepsilon > 0$ and N are given, for k large enough $\pi_k^{\otimes N}$ is concentrated on the set $\mathcal{C}_\varepsilon(N)$ defined by

$$\sum_{1 \le k \le N} c(x_i, y_i) \le \sum_{1 \le k \le N} c(x_i, y_{i+1}) + \varepsilon.$$

Since this is a closed set, the same is true for $\pi^{\otimes N}$, and then by letting $\varepsilon \to 0$ we see that $\pi^{\otimes N}$ is concentrated on the set $\mathcal{C}(N)$ defined by

$$\sum_{1 \le k \le N} c(x_i, y_i) \le \sum_{1 \le k \le N} c(x_i, y_{i+1}).$$

So the support of π is c-cyclically monotone, as desired.

Now assume that $\liminf_{k \to \infty} \int c_k \, d\pi_k < +\infty$. Then

$$\int c \, d\pi \le \liminf_{k \to \infty} \int c_k \, d\pi_k < +\infty.$$

In particular, $C(\mu, \nu) < +\infty$; so Theorem 5.9 applies and guarantees the optimality of π. □

Theorem 5.18 admits the following corollary about the stability of *transport maps*.

Corollary 5.20 (Stability of the transport map). *With the same assumptions and notation as in Theorem 5.18, further assume that $\mathcal{Y}$ is a smooth Riemannian manifold, that there exist measurable maps $T_k, T : \mathcal{X} \to \mathcal{Y}$ such that*

$$\pi_k = (\mathrm{Id}\,, T_k)_\# \mu_k; \qquad \pi = (\mathrm{Id}\,, T)_\# \mu;$$

and that π is the unique optimal transference plan between μ and ν. Then

$$\forall \varepsilon > 0 \qquad \mu_k \Big[\big\{ d(T_k, T) > \varepsilon \big\} \Big] \xrightarrow[k \to \infty]{} 0,$$

where d denotes the distance in $\mathcal{Y}$. In particular, if $\mu_k = \mu$ for all k, then T_k converges to T in μ-probability.

Proof of Corollary 5.20. By Theorem 5.18 and uniqueness of the optimal coupling between μ and ν, we know that $\pi_k = (\mathrm{Id}\,, T_k)_\# \mu_k$ converges weakly to $\pi = (\mathrm{Id}\,, T)_\# \mu$.

Let $\delta > 0$ be given, and let $y_0 \in \mathcal{Y}$ be arbitray. There is a compact set $K \subset \mathcal{X}$ such that $\mu[X \setminus K] \leq \delta$. Since $\mathcal{Y}$ is a Riemannian manifold, there is $R > 0$ such that $\mu[\{d(y_0, T(x)) > R\}] = \nu[\{d(y_0, y) > R\}] \leq \delta$. Let $T_\delta(x)$ be defined by $T_\delta(x) = T(x)$ if $x \in K$ and $d(y_0, T(x)) \leq R$; and $T_\delta(x) = y_0$ otherwise. The map T_δ takes values in a compact subset of a Riemannian manifold, which can be covered by finitely many diffeomorphisms valued in some subset of $\mathbb{R}^m$. By using a partition of unity and the usual Lusin theorem for functions valued in $\mathbb{R}^m$, we can construct a *continuous* function $\widetilde{T}_\delta$, constant outside of a compact set, such that $\mu[\{\widetilde{T}_\delta \neq T_\delta\}] \leq \delta$; and as a consequence $\mu[\{\widetilde{T}_\delta \neq T\}] \leq 3\delta$; Then

$$\pi\Big[\big\{(x,y);\ y \neq \widetilde{T}_\delta(x)\big\}\Big] \leq \pi\Big[\big\{(x,y);\ y \neq T(x)\big\}\Big] + 3\delta.$$

Let then $O_\delta := \{(x,y) \in \mathcal{X} \times \mathcal{Y};\ d(y, \widetilde{T}(x)) < \delta\}$. Since $\widetilde{T}$ is continuous, O_δ is open, so

$$1 = \pi[O_\delta] \leq \liminf_{\delta \to 0} \pi_k[O_\delta].$$

So $\pi_k[O_\delta] \to 1$, which means

$$\mu_k\Big[\big\{x \in \mathcal{X};\ d(T_k(x), \widetilde{T}(x)) < \delta\big\}\Big] \xrightarrow[k\to\infty]{} 1.$$

This concludes the argument. □

Application: Dual formulation of transport inequalities

Let

$$C(\mu,\nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int c\, d\pi \tag{5.28}$$

stand for the value of the optimal transport cost of transport between μ and ν.

If ν is a given reference measure, inequalities of the form

$$\forall \mu \in P(\mathcal{X}), \qquad \inf_{\pi \in \Pi(\mu,\nu)} \int c\, d\pi \leq F(\mu)$$

arise in several branches of mathematics; some of them will be studied in Chapter 22. It is useful to know that if F is a convex function of μ, then there is a nice dual reformulation of these inequalities in terms of the Legendre transform of F. This is the content of the following theorem.

Theorem 5.21 (Dual transport inequalities). *Let $\mathcal{X}$, $\mathcal{Y}$ be two Polish spaces, and ν a given probability measure on $\mathcal{Y}$. Let $F : P(\mathcal{X}) \to \mathbb{R} \cup \{+\infty\}$ be a convex lower semi-continuous function on $P(\mathcal{X})$, and Λ its Legendre transform on $C_b(\mathcal{X})$; more explicitly, it is assumed that*

$$\begin{cases} \forall \mu \in P(\mathcal{X}), \quad F(\mu) = \sup_{\varphi \in C_b(\mathcal{X})} \displaystyle\int_{\mathcal{X}} \varphi\, d\mu - \Lambda(\varphi) \\ \forall \varphi \in C_b(\mathcal{X}), \quad \Lambda(\varphi) = \sup_{\mu \in P(\mathcal{X})} \displaystyle\int_{\mathcal{X}} \varphi\, d\mu - F(\mu). \end{cases} \tag{5.29}$$

Let further $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ be a lower semi-continuous cost function. Then, the following two statements are equivalent:

(i) $\forall \mu \in P(\mathcal{X}), \qquad C(\mu, \nu) \le F(\mu)$;

(ii) $\forall \phi \in C_b(\mathcal{Y}), \qquad \Lambda\left(\int_{\mathcal{Y}} \phi \, d\nu \; - \; \phi^c\right) \le 0$, *where* $\phi^c(x) := \sup_y [\phi(y) - c(x,y)]$.

Moreover, if $\Phi : \mathbb{R} \to \mathbb{R}$ is an increasing convex function, and Φ^* is its Legendre transform, then the following two statements are equivalent:

(i') $\forall \mu \in P(\mathcal{X}), \qquad \Phi(C(\mu, \nu)) \le F(\mu)$;

(ii') $\forall \phi \in C_b(\mathcal{Y}), \; \forall t \ge 0 \qquad \Lambda\left(t \int_{\mathcal{Y}} \phi \, d\nu \; - \; t\phi^c - \Phi^*(t)\right) \le 0.$

Remark 5.22. The writing in (ii) or (ii') is not very rigorous since Λ is a priori defined on the set of bounded continuous functions, and ϕ^c might not belong to that set. (It is clear that ϕ^c is bounded from above, but this is all that can be said.) However, from (5.29) Λ is a nondecreasing function of φ, so there is in practice no problem to extend it to a more general class of measurable functions. In any case, the correct way to interpret the left-hand side in (ii) is

$$\Lambda\left(\int_{\mathcal{Y}} \phi \, d\nu \; - \; \phi^c\right) = \sup_{\psi \ge \phi^c} \Lambda\left(\int_{\mathcal{Y}} \phi \, d\nu \; - \; \psi\right),$$

where ψ in the infimum is assumed to be bounded continuous.

Remark 5.23. One may simplify (ii') by taking the supremum over t; since Λ is nonincreasing, the result is

$$\Lambda\left(\Phi\left(\int_{\mathcal{Y}} \phi \, d\nu \; - \; \phi^c\right)\right) \le 0. \tag{5.30}$$

(This shows in particular that the equivalence (i) $\Leftrightarrow$ (ii) is a particular case of the equivalence (i') $\Leftrightarrow$ (ii').) However, in certain situations it is better to use the inequality (ii') rather than (5.30); see for instance Proposition 22.2.

Example 5.24. The most famous example of inequality of the type of (i) is when $F(\mu)$ is the Kullback information with respect to μ, that is $F(\mu) = H_\nu(\mu) = \int \rho \log \rho \, d\nu$, where ρ is the density of μ with respect to the reference probability measure ν; and by convention $F(\mu) = +\infty$ if μ is not absolutely continuous with respect to ν. Then one has the explicit formula

$$\Lambda(\varphi) = \log\left(\int e^{\varphi} \, d\nu\right).$$

So the two functional inequalities

$$\forall \mu \quad C(\mu, \nu) \le H_\nu(\mu)$$

and

$$\forall \phi \quad \int e^{-\phi^c} \, d\nu \le e^{-\int \phi \, d\nu}$$

are equivalent.

Proof of Theorem 5.21. First assume that (i) is satisfied. Then for all $\psi \geq \phi^c$,

$$\begin{aligned}\Lambda\left(\int_{\mathcal{Y}} \phi\, d\nu \,-\, \psi\right) &= \sup_{\mu\in P(\mathcal{X})} \left\{\int_{\mathcal{X}} \left(\int_{\mathcal{Y}} \phi\, d\nu \,-\, \psi\right) d\mu - F(\mu)\right\} \\ &= \sup_{\mu\in P(\mathcal{X})} \left\{\int_{\mathcal{Y}} \phi\, d\nu - \int_{\mathcal{X}} \psi\, d\mu - F(\mu)\right\} \\ &\leq \sup_{\mu\in P(\mathcal{X})} \Big[C(\mu,\nu) - F(\mu)\Big] \leq 0,\end{aligned}$$

where the most easy part of Theorem 5.9 (that is, inequality (5.4)) was used to go from the but-to-last line to the last one. Then (ii) follows upon taking the supremum over ψ.

Conversely, assume that (ii) is satisfied. Then, for any pair $(\psi, \phi) \in C_b(\mathcal{X}) \times C_b(\mathcal{Y})$ one has, by (5.29),

$$\int_{\mathcal{Y}} \phi\, d\nu - \int_{\mathcal{X}} \psi\, d\mu = \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} \phi\, d\nu - \psi\right) d\mu \leq \Lambda\left(\int_{\mathcal{Y}} \phi\, d\nu - \psi\right) \,+\, F(\mu).$$

Taking the supremum over all $\psi \geq \phi^c$ yields

$$\int_{\mathcal{Y}} \phi\, d\nu - \int_{\mathcal{X}} \phi^c\, d\mu \leq \Lambda\left(\int_{\mathcal{Y}} \phi\, d\nu - \phi^c\right) \,+\, F(\mu).$$

By assumption, the first term in the right-hand side is always nonpositive; so in fact

$$\int_{\mathcal{Y}} \phi\, d\nu - \int_{\mathcal{X}} \phi^c\, d\mu \leq F(\mu).$$

Then (i) follows upon taking the supremum over $\phi \in C_b(\mathcal{Y})$ and applying Theorem 5.9 (i).

Now let us consider the equivalence between (i') and (ii'). Since Φ is increasing on $\mathbb{R}$, it is easy to check that its Legendre transform Φ^* takes the value $+\infty$ on $(-\infty, 0)$. So the Legendre inversion formula becomes

$$\forall r \in \mathbb{R}, \quad \Phi(r) = \sup_{t\in\mathbb{R}_+} \big[rt - \Phi^*(t)\big].$$

(The important thing is that the supremum is over $\mathbb{R}_+$ and not $\mathbb{R}$.)

If (i') is satisfied, then for all $\phi \in C_b(\mathcal{X})$, for all $\psi \geq \phi^c$ and for all $t \in \mathbb{R}_+$,

$$\begin{aligned}\Lambda\left(t\int \phi\, d\nu - t\psi - \Phi^*(t)\right) &= \sup_{\mu\in P(\mathcal{X})} \left[\left(t\int \phi\, d\nu - t\psi - \Phi^*(t)\right) d\mu - F(\mu)\right] \\ &= \sup_{\mu\in P(\mathcal{X})} \left[t\left(\int_{\mathcal{X}} \phi\, d\nu - \int \psi\, d\mu\right) - \Phi^*(t) - F(\mu)\right] \\ &\leq \sup_{\mu\in P(\mathcal{X})} \Big[tC(\mu,\nu) - \Phi^*(t) - F(\mu)\Big] \\ &\leq \sup_{\mu\in P(\mathcal{X})} \Big[\Phi(C(\mu,\nu)) - F(\mu)\Big] \leq 0,\end{aligned}$$

where the inequality $tr \leq \Phi(r) + \Phi^*(t)$ was used.

On the other hand, if (ii') is satisfied, then for all $(\phi, \psi) \in C_b(\mathcal{X}) \times C_b(\mathcal{Y})$ and $t \geq 0$,

$$t \int \phi \, d\nu - t \int \psi \, d\mu - \Phi^*(t) = \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} t\phi \, d\nu - t\psi - \Phi^*(t) \right) d\mu$$
$$\leq \Lambda \left(t \int \phi \, d\nu - t\psi - \Phi^*(t) \right) + F(\mu);$$

then by taking the supremum over $\psi \geq \phi^c$ one obtains

$$t\, C(\mu, \nu) - \Phi^*(t) \leq \Lambda \left(t \int \phi \, d\nu - t\phi^c - \Phi^*(t) \right) + F(\mu)$$
$$\leq F(\mu);$$

then (i') follows by taking the supremum over $t \geq 0$. □

Application: Solvability of the Monge problem

As a last application of Theorem 5.9, I shall now present the criterion which is used in the large majority of proofs of existence of a deterministic optimal coupling (or Monge transport).

Theorem 5.25 (Criterion for solvability of the Monge problem). *Let $\mathcal{X}$ and $\mathcal{Y}$ be two Polish spaces, and let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a lower semi-continuous cost function. Let $\mu \in P(\mathcal{X})$ and $\nu \in P(\mathcal{Y})$ be two Borel probability measures, and let $C(\mu, \nu)$ the optimal total transport cost between μ and ν. If*

(i) $C(\mu, \nu) < +\infty$;

(ii) For any c-convex function $\psi : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$, the set of $x \in \mathcal{X}$ such that $\partial_c \psi(x)$ contains more than one element is μ-negligible;

Then, there is a unique (in law) optimal coupling (X, Y) of (μ, ν), and it is deterministic. It is characterized (among all possible couplings) by the existence of a c-convex function ψ such that, almost surely, $Y \in \partial_c \psi(X)$. In particular, the Monge problem with initial measure μ and final measure ν admits a unique solution.

Proof of Theorem 5.25. The argument is almost obvious. By Theorem 5.9(ii), there is a c-convex function ψ, and a measurable set $\Gamma \subset \partial_c \psi$ such that any optimal plan π is concentrated on Γ. By assumption there is a Borel set Z such that $\mu[Z] = 0$ and $\partial_c \psi(x)$ contains at most one element if $x \notin Z$. So for any $x \in \mathrm{proj}_{\mathcal{X}}(\Gamma) \setminus Z$, there is exactly one $y \in \mathcal{Y}$ such that $(x, y) \in \Gamma$, and we can then define $T(x) = y$.

Let now π be any optimal coupling. As I just said, it has to be concentrated on Γ; and since $Z \times \mathcal{Y}$ is π-negligible, it is also concentrated on $\Gamma \setminus (Z \times \mathcal{Y})$, which is precisely the set of all couples of the form $(x, T(x))$, or the graph of T. It follows that π is the Monge transport associated with the map T.

The argument above is in fact a bit sloppy, since I did not check the measurability of T. So I shall show below how to slightly modify the construction of T to make sure that it is measurable. The reader who does not want to bother about measurability issues can skip the rest of the proof.

Let $(K_\ell)_{\ell \in \mathbb{N}}$ be a nondecreasing sequence of compact sets, all of them included in $\Gamma \setminus (Z \times \mathcal{Y})$, such that $\pi[\cup K_\ell] = 1$. (The family (K_ℓ) exists because π, just as any Borel measure on a Polish space, is regular.) If ℓ is given, then for any x lying in the compact set $J_\ell := \mathrm{proj}_{\mathcal{X}}(K_\ell)$ we can define $T_\ell(x)$ as the unique y such that $(x, y) \in K_\ell$. Then we can define T on $\cup J_\ell$ by the requirement that for each ℓ, T restricts to T_ℓ on J_ℓ. Because K_ℓ is

compact, it is trivial to check the continuity of the map T_ℓ on J_ℓ; and then it follows that T is a Borel map. Even if it is not defined on the whole of $\Gamma \setminus (Z \times \mathcal{Y})$, it is still defined on a set of full μ-measure, so the proof can be concluded just as before. □

Bibliographical Notes

There are many ways to state the Kantorovich duality, and even more ways to prove it. There are also several economic interpretations, that belong to folklore. The one which I formulated in this chapter is a variant of one that I learnt from Caffarelli. Related economic interpretations underlie some algorithms, such as the fascinating "auction algorithm" developed by Bertsekas (see [50, Chapter 7], or the various surveys written by Bertsekas on the subject).

A common convention consists in taking the pair $(-\psi, \phi)$ as the unknown. (The latter pair was denoted (φ, ψ) in [365, Chapter 1], which will probably upset the reader.) This has the advantage to make some formulas more symmetric: The c-transform becomes $\varphi^c(y) = \inf_x [c(x,y) - \varphi(x)]$, and then $\psi^c(x) = \inf_y [c(x,y) - \psi(y)]$, so this is the same formula going back and forth between functions of x and functions of y, upon exchange of x and y. Since in general $\mathcal{X}$ and $\mathcal{Y}$ have nothing in common, this symmetry is essentially cosmetic. The conventions used in this chapter lead to a somewhat natural "economic" interpretation, and will also lend themselves better to a time-dependent treatment. Moreover, they also agree with the conventions used in the Aubry–Mather–Fathi weak KAM theory [160, 47]. It might be good to make the link more explicit. In weak KAM theory, $\mathcal{X} = \mathcal{Y}$ is a Riemannian manifold M, a Lagrangian cost function is given on the tangent bundle TM, and $c(x,y)$ is a continuous cost function defined from the dynamics, as the minimum action that one should spend to go from x to y (as later in Chapter 7). Since in general $c(x,x) \neq 0$, it is not absurd to consider the optimal transport cost $C(\mu,\mu)$ between a measure μ and itself. If M is compact, it is easy to show that there exists a $\overline{\mu}$ that minimizes $C(\mu,\mu)$. To the optimal transport problem between $\overline{\mu}$ and $\overline{\mu}$, Theorem 5.9 associates a minimal and a maximal closed c-cyclically monotone sets, respectively $\Gamma_{\min}$ and $\Gamma_{\max} \subset M \times M$. These sets can be identified with subsets of TM via the embedding (initial position, final position) $\longmapsto$ (initial position, initial velocity). Under that identification, $\Gamma_{\min}$ and $\Gamma_{\max}$ are called respectively the Mather and Aubry sets; they carry valuable information about the underlying dynamics. For mnemonic purposes, to recall which is which, the reader might use the resemblance of the name "Mather" with the word "measure". (The Mather set is the one cooked up from the supports of the probability measures.)

In the particular case when $c(x,y) = |x-y|^2/2$ in Euclidean space, then it is customary to expand this as $|x|^2/2 - x \cdot y + |y|^2/2$, and change unknowns by including $|x|^2/2$ and $|y|^2/2$ into ψ and ϕ respectively, then change signs and reduce to the cost function $x \cdot y$, which is the one appearing naturally in the Legendre duality of convex functions. This is explained carefully in [365, Chapter 2], where reminders and references about the theory of convex functions in $\mathbb{R}^n$ are also provided.

The duality theorem was proven by Kantorovich and Rubinstein in the particular case when $c(x,y)$ is the distance on a compact space; then the statement was generalized by Dudley [145, 147], who also showed that even the completeness can be dispended with (the proof in the first reference contains a gap which was filled by de Acosta [128, Appendix B], following an idea suggested by Dudley in Saint-Flour, 25 years ago!). Rüschendorf [173, 317], Fernique [164], Szulga [345], Kellerer [226], and probably others, contributed to the problem. Another short proof was also communicated to me by Feyel.

Modern treatments most often use variants of the Hahn-Banach theorem, see for instance [306, 365]. The proof presented in [365, Theorem 1] proves the duality when $\mathcal{X}, \mathcal{Y}$ are compact, then treats the general case by an approximation argument; this is somewhat tricky but has the advantage to avoid the general version of the axiom of choice, since it uses the Hahn-Banach theorem only in the separable space $C(K)$, where K is compact.

Ramachandran and Rüschendorf [309, 310] investigated the Kantorovich duality out of the setting of Polish spaces, and found out a necessary and sufficient condition, called "perfection", for its validity.

In the case when the cost function is a distance, the optimal transport problem coincides with the Kantorovich transshipment problem, for which more duality theorems are available, and a vast literature has been written; see [306] for results and references.

Around the mid-eighties, it was understood that the study of the dual problem, and in particular the existence of a maximizer, could lead to precious qualitative information about the solution of the Monge–Kantorovich problem. This point of view was emphasized by Knott and Smith [229], Brenier [69, 72], Rachev and Rüschendorf [321, 306], Gangbo [176], Gangbo and McCann [178, 179], McCann [269] and others. Then Ambrosio and Pratelli proved the existence of a maximizing pair under the precise conditions 5.12; see [17, Theorem 3.2]. Under adequate assumptions, one can also prove the existence of a maximizer for the dual problem by direct arguments which do not use the original problem (see for instance [365, Chapter 2]).

The notion of c-convexity, as a generalization of the usual notion of convexity, was studied by several authors, in particular Rüschendorf [321].

For the proof of Theorem 5.9, I borrowed from McCann [266] the idea of recovering c-cyclical monotonicity from approximation by combinations of Dirac masses; from Rüschendorf [320] the method used to construct ψ from Γ; from Schachermayer and Teichmann [328] the clever truncation procedure used in the proof of Part (ii); apart from that the general scheme of proof is more or less the one used by Ambrosio and Pratelli [17], and Ambrosio, Gigli and Savaré [15]. On the whole, the proof avoids not only the use of the axiom of choice, but also any version of the Hahn-Banach theorem; and it also leads to the best known results. In my opinion this does compensate for its being somewhat tricky.

About the proof of the Kantorovich duality, it is interesting to notice that "duality for somebody implies duality for everybody" (a rule which is true in other branches of analysis): In the present case, constructing one particular cyclically monotone transference plan allows one to prove the duality, which leads to the conclusion that *all* transference plans should be cyclically monotone. By the way, the latter statement could also be proven directly with the help of a bit of measure theoretical abstract nonsense, see e.g. [178].

The use of the law of large numbers for empirical measures might be natural for a probabilistic audience, but one should not forget that this is a subtle result: For any bounded continuous test function, the usual law of large numbers yields convergence out of a negligible set, but then one has to find a negligible set that works for *all* bounded continuous test functions. Dudley [147, Theorem 11.4.1] proves this for general separable metric spaces, giving credit to Varadarajan for this theorem. In the community of dynamical systems, these results are known as part of the so-called Krylov-Bogoljubov theory, in relation with ergodic theorems; see e.g. Oxtoby [295] for a compact space.

The equivalence between the properties of optimality (of a transference plan) and cyclical monotonicity, for quite general cost functions and probability measures, was a widely open problem until recently; it was explicitly listed as Open Problem 2.25 in [365] for a quadratic cost function in $\mathbb{R}^n$. The current state of the art is that

- the equivalence is *false* for a general lower semi-continuous function with possibly infinite values, as shown by a clever counterexample of Ambrosio and Pratelli [17];

- the equivalence is true for a *continuous* cost function with possibly infinite values, as shown by Pratelli [302];

- the equivalence is true for a *real-valued* lower semi-continuous cost function, as shown by Schachermayer and Teichmann [328]; this is the result that I chose to present in these notes.

Schachermayer and Teichmann gave a nice interpretation of the Ambrosio–Pratelli counterexample and suggested that the correct notion in the whole business is not cyclical monotonicity, but a variant which they named "strong cyclical monotonicity condition" [328].

As I am writing these notes, it seems that the final resolution of this equivalence issue will soon be available, but at the price of a journey into the very guts of measure theory. The following construction was explained to me by Bianchini. (Skip if you fear abstraction.) Let c be an arbitrary lower semi-continuous cost function with possibly infinite values, and let π be a cyclically monotone plan. Let Γ be a cyclically monotone set with $\pi[\Gamma] = 1$. Define an equivalence relation $\mathcal{R}$ on Γ as follows: $(x, y) \sim (x', y')$ if there is a finite number of couples (x_k, y_k), $0 \leq k \leq N$, such that: $(x, y) = (x_0, y_0)$; either $c(x_0, y_1) < +\infty$ or $c(x_1, y_0) < +\infty$; $(x_1, y_1) \in \Gamma$; either $c(x_1, y_2) < +\infty$ or $c(x_2, y_1) < +\infty$; etc. until $(x_N, y_N) = (x', y')$. The relation $\mathcal{R}$ divides Γ into equivalence classes $(\Gamma_\alpha)_{\alpha \in \Gamma/\mathcal{R}}$. Let p be the map which to a point x associates its equivalence class $\overline{x}$. The set $\Gamma/\mathcal{R}$ in general has no topological or measurable structure, but we can equip it with the largest σ-algebra making p measurable. On $\Gamma \times (\Gamma/\mathcal{R})$ introduce the product σ-algebra. If now the *graph* of p is measurable for this product measure, then π should be optimal in the Monge–Kantorovich problem.

In most applications, the cost function is continuous, and often rather simple. However, it is sometimes useful to consider cost functions that achieve the value $+\infty$, as in the "secondary variational problem" considered by Ambrosio and Pratelli [17] or by Bernard and Buffoni [48]. Such is also the case for the optimal transport in Wiener space considered by Feyel and Üstünel [165, 166, 168, 167], for which the cost function $c(x, y)$ is the square norm of $x - y$ in the Cameron–Martin space (so it is $+\infty$ if $x - y$ does not belong to that space). In this setting, optimizers in the dual problem can be constructed via finite-dimensional approximations, but it is not known whether there is a more direct construction by c-monotonicity.

When condition (5.9) (or its weakened version (5.12)) is relaxed, it is not clear in general that the dual Kantorovich problem admits a maximizing pair. Yet this is true for instance in the case of optimal transport in Wiener space; this is an indication that condition (5.12) might not be the "correct" one, although at present no better general condition is known.

Theorem 5.16 was inspired by a recent work of Fathi and Figalli [161], in which a restriction procedure is used to solve the Monge problem for certain cost functions arising from Lagrangian dynamics in unbounded phase spaces; see Theorem 10.26 below for more information.

Theorem 5.21 appears in a more or less explicit form in various works, especially for the important example $F(\mu) = \int \rho \log \rho \, d\nu$, where ρ stands for the density of μ with respect to ν (and $F(\mu) = +\infty$ if μ is singular with respect to ν); see for instance [307].

Finally, a few words about basic measure-theoretical tools. The regularity of Borel measures on Polish spaces is proven in [147, p. 225]. Lusin's theorem states the following: If $F : \mathcal{X} \to \mathbb{R}$ is a measurable function defined on a locally compact measure space with finite mass, then for any $\varepsilon > 0$ there is a continuous function $\widetilde{F} : \mathcal{X} \to \mathbb{R}$ with compact

support, such that F and $\widetilde{F}$ coincide up to a set of measure at most ε. This statement easily extends to functions valued in $\mathbb{R}^m$.

6

The Wasserstein distances

Assume, as before, that you are in charge of the transport of goods between producers and consumers, whose respective spatial distributions are modelled by probability measures. The more producers and consumers are far away from each other, the more difficult will be your job, and you would like to summarize the degree of difficulty with just one quantity. For that purpose it is natural to consider, as in (5.28), the **optimal transport cost** between the two measures:

$$C(\mu,\nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int c(x,y)\, d\pi(x,y), \tag{6.1}$$

where $c(x,y)$ is the cost for transporting one unit of mass from x to y. Here we do not care about the shape of the optimizer but only in the *value* of this optimal cost.

One can think of (6.1) as a kind of distance between μ and ν, but in general it does not, strictly speaking, satisfy the axioms of a distance function. However, when the cost is defined in terms of a distance, it is easy to cook up a distance from (6.1).

Definition 6.1 (Wasserstein distances). *Let $(\mathcal{X}, d)$ be a metric space, and let $p \in [1, +\infty)$. For any two probability measures μ, ν on $\mathcal{X}$, the Wasserstein distance of order p between μ and ν is defined by the formula*

$$\begin{aligned} W_p(\mu,\nu) &= \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}} d(x,y)^p \, d\pi(x,y) \right)^{1/p} \\ &= \inf \left\{ \Big[\mathbb{E}\, d(X,Y)^p \Big]^{\frac{1}{p}}, \quad \text{law}\,(X) = \mu, \quad \text{law}\,(Y) = \nu \right\}. \end{aligned} \tag{6.2}$$

Example 6.2. $W_p(\delta_x, \delta_y) = d(x,y)$. In this example, the distance does not depend on p; but this is not the rule.

At the present level of generality, W_p is not a distance in the strict sense, because it might take the value $+\infty$; but otherwise it does satisfy the axioms of a distance, as will be checked right now.

Proof that W_p satisfies the axioms of a distance. First, it is clear that $W_p(\mu,\nu) = W_p(\nu,\mu)$.

Next, let μ_1, μ_2 and μ_3 be three probability measures on $\mathcal{X}$, and let (X_1, X_2) be an optimal coupling of (μ_1, μ_2) (for the cost function $c = d^p$), and (X_2, X_3) an optimal coupling of (μ_2, μ_3). By the Gluing Lemma (recalled in Chapter 1), there exist random variables (X_1', X_2', X_3') with $\text{law}\,(X_1', X_2') = \text{law}\,(X_1, X_2)$ and $\text{law}\,(X_2', X_3') = \text{law}\,(X_2, X_3)$. In particular, (X_1', X_3') is a coupling of (μ_1, μ_3), so

$$W_p(\mu_1,\mu_3) \le \left(\mathbb{E}\, d(X_1',X_3')^p\right)^{\frac{1}{p}} \le \left(\mathbb{E}\left(d(X_1',X_2') + d(X_2',X_3')\right)^p\right)^{\frac{1}{p}}$$
$$\le \left(\mathbb{E}\, d(X_1',X_2')^p\right)^{\frac{1}{p}} + \left(\mathbb{E}\, d(X_2',X_3')^p\right)^{\frac{1}{p}}$$
$$= W_p(\mu_1,\mu_2) + W_p(\mu_2,\mu_3),$$

where the inequality leading to the second line is an application of the Minkowski inequality in $L^p(\mathbb{P})$, and the last equality follows from the fact that (X_1', X_2') and (X_2', X_3') are optimal couplings. So W_p satisfies the triangular inequality.

Finally, assume that $W_p(\mu,\nu) = 0$; then there exists a transference plan which is entirely concentrated on the diagonal $(y = x)$ in $\mathcal{X} \times \mathcal{X}$. So $\nu = \mathrm{Id}_{\,\#}\mu = \mu$. □

To complete the construction it is natural to restrict W_p to a subset of $P(\mathcal{X}) \times P(\mathcal{X})$ on which it takes finite values.

Definition 6.3 (Wasserstein space). *With the same conventions as in Definition 6.1, the Wasserstein space of order p is defined as*

$$P_p(\mathcal{X}) := \left\{\mu \in P(\mathcal{X}); \qquad \int_{\mathcal{X}} d(x_0,x)^p\, d\mu(x) < +\infty\right\},$$

where $x_0 \in \mathcal{X}$ is arbitrary. This space does not depend on the choice of the point x_0. Then W_p defines a (finite) distance on $P_p(\mathcal{X})$.

In words, the Wasserstein space is the space of probability measures which have a *finite moment of order p*. In these notes, it will always be equipped with the metric W_p.

Proof that W_p is finite on P_p. Let π be a transference plan between two elements μ and ν in $P_p(\mathcal{X})$. Then the inequality

$$d(x,y)^p \le 2^{p-1}\left[d(x,x_0)^p + d(x_0,y)^p\right]$$

shows that $d(x,y)^p$ is $\pi(dx\,dy)$-integrable as soon as $d(\cdot,x_0)^p$ is μ-integrable and $d(x_0,\cdot)^p$ is ν-integrable. □

Remark 6.4. The combination of Theorem 5.9 and Particular Case 5.4 leads to the useful **duality formula for the Kantorovich–Rubinstein distance**: For any $\mu, \nu \in P_1(\mathcal{X})$,

$$W_1(\mu,\nu) = \sup_{\|\psi\|_{\mathrm{Lip}} \le 1} \left\{\int_{\mathcal{X}} \psi\, d\mu - \int_{\mathcal{X}} \psi\, d\nu\right\}. \tag{6.3}$$

Remark 6.5. A simple application of Hölder's inequality shows that

$$p \le q \Longrightarrow W_p \le W_q. \tag{6.4}$$

In particular, the Wasserstein distance of order 1, W_1, is the weakest of all. The most useful exponents in the Wasserstein distances are $p = 1$ (**Kantorovich–Rubinstein distance**) and $p = 2$. As a general rule, the W_1 distance is more flexible and convenient for getting bounds, while the W_2 distance better reflects the geometric features of the problem, and is better adapted when there is more structure; it also scales better with the dimension. Results in W_2 distance are usually stronger, and more difficult to establish, than results in W_1 distance.

Convergence in Wasserstein sense

Here is a characterization of convergence in the Wasserstein space.

Definition 6.6 (Weak convergence in P_p). *Let $(\mathcal{X}, d)$ be a Polish space, and $p \in [1, \infty)$. Let $(\mu_k)_{k\in\mathbb{N}}$ and μ be elements of $P_p(\mathcal{X})$. Then (μ_k) is said to converge weakly in $P_p(\mathcal{X})$ if any one of the following equivalent properties is satisfied for some (and then any) $x_0 \in \mathcal{X}$:*

(i) μ_k converges weakly to μ and $\displaystyle\int d(x_0, x)^p \, d\mu_k(x) \longrightarrow \int d(x_0, x)^p \, d\mu(x)$;

(ii) μ_k converges weakly to μ and $\displaystyle\limsup_{k\to\infty} \int d(x_0, x)^p \, d\mu_k(x) \le \int d(x_0, x)^p \, d\mu(x)$;

(iii) μ_k converges weakly to μ and $\displaystyle\lim_{R\to\infty} \limsup_{k\to\infty} \int_{d(x_0,x)\ge R} d(x_0, x)^p \, d\mu_k(x) = 0$.

(iv) For all continuous functions φ with $|\varphi(x)| \le C\big(1 + d(x_0, x)^p\big)$, $C \in \mathbb{R}$, one has

$$\int \varphi(x) \, d\mu_k(x) \longrightarrow \int \varphi(x) \, d\mu(x).$$

Theorem 6.7 (W_p metrizes P_p). *Let $(\mathcal{X}, d)$ be a Polish space, and $p \in [1, \infty)$; then the Wasserstein distance W_p metrizes the weak convergence in $P_p(\mathcal{X})$. In other words, if $(\mu_k)_{k\in\mathbb{N}}$ is a sequence of measures in $P_p(\mathcal{X})$ and μ is another measure in $P(\mathcal{X})$, then the statements*

$$\mu_k \text{ converges weakly in } P_p(\mathcal{X}) \text{ to } \mu$$

and

$$W_p(\mu_k, \mu) \longrightarrow 0$$

are equivalent.

Here are two immediate corollaries of this theorem (the first one results from the triangular inequality):

Corollary 6.8 (Continuity of W_p). *If $(\mathcal{X}, d)$ is a Polish space, and $p \in [1, \infty)$, then W_p is continuous on $P_p(\mathcal{X})$. More explicitly, if μ_k (resp. ν_k) converges to μ (resp. ν) weakly in $P_p(\mathcal{X})$ as $k \to \infty$, then*

$$W_p(\mu_k, \nu_k) \longrightarrow W_p(\mu, \nu).$$

Remark 6.9. On the contrary, if these convergences are only usual weak convergences, then one can only conclude that $W_p(\mu, \nu) \le \liminf W_p(\mu_k, \nu_k)$: the Wasserstein distance is lower semi-continuous on $P(\mathcal{X})$ (just like the optimal transport cost C, for any lower semi-continuous cost function c; recall the proof of Theorem 4.1).

Corollary 6.10 (metrizability of the weak topology). *Let $(\mathcal{X}, d)$ be a Polish space. If $\widetilde{d}$ is a bounded distance inducing the same topology as d (e.g. $\widetilde{d} = d/(1 + d)$), then the convergence in Wasserstein sense for the distance $\widetilde{d}$ is equivalent to the usual weak convergence of probability measures in $P(\mathcal{X})$.*

Before starting the proof of Theorem 6.7, it will be good to make some more comments. The short version of that theorem is that *Wasserstein distances metrize weak convergence.* This sounds good, but after all, there are many ways to metrize weak convergence. Here below are some of the most popular ones, defined either in terms of measures μ, ν, or in terms of random variables X, Y with $\text{law}\,(X) = \mu$, $\text{law}\,(Y) = \nu$:

- the **Lévy-Prokhorov distance**:

$$d_P(\mu,\nu) = \inf\Big\{\varepsilon > 0;\ \inf \mathbb{P}\left[d(X,Y) > \varepsilon\right] \le \varepsilon\Big\}; \tag{6.5}$$

- the **bounded Lipschitz distance** (also called the Fortet-Mourier distance):

$$d_{bL}(\mu,\nu) = \sup\left\{\int \varphi\, d\mu - \int \varphi\, d\nu;\quad \|\varphi\|_\infty + \|\varphi\|_{\mathrm{Lip}} \le 1\right\}; \tag{6.6}$$

- the **weak-$*$ distance** (on a locally compact metric space):

$$d_{w*}(\mu,\nu) = \sum_{k=0}^{\infty} 2^{-k} \left|\int \varphi_k\, d\mu - \int \varphi_k\, d\nu\right|, \tag{6.7}$$

  where $(\varphi_k)_{k\in\mathbb{N}}$ is a dense sequence in $C_0(\mathcal{X})$;

- the **Toscani metric** (on $P_2(\mathbb{R}^n)$)

$$d_T(\mu,\nu) = \sup_{\xi\in\mathbb{R}^n\setminus\{0\}} \left(\frac{\left|\int e^{-ix\cdot\xi}\, d\mu(x) \ - \ \int e^{-ix\cdot\xi}\, d\nu(x)\right|}{|\xi|^2}\right) \qquad (i^2 = -1) \tag{6.8}$$

  (Here I implicitly assume that μ, ν have the same mean, otherwise $d_T(\mu,\nu)$ might be infinite; one can also introduce variants of d_T by changing the exponent 2 in the denominator.)

So why bother with Wasserstein distances? There are several answers to that question:

1. Wasserstein distances are rather strong, especially in the way they take care of large distances in $\mathcal{X}$; this is a definite advantage over, for instance, the weak-$*$ distance (which in practice is so weak that I advice the reader to never use it). It is not so difficult to combine an information of convergence in Wasserstein distance with some smoothness bound, in order to get convergence in stronger distances.
2. The definition of Wasserstein distances makes them convenient to use in many problems where optimal transport is naturally involved; as in many problems coming from partial differential equations. This is an advantage over, for instance, the Toscani metric which is a bit more artificial.
3. The Wasserstein distances have a rich duality; this is especially useful for $p = 1$, in view of (6.3) (compare with the definition of the bounded Lipschitz distance). Passing back and forth from the original to the dual definition is often technically convenient.
4. Wasserstein distances are defined by an infimum, which from a technical point of view often makes them relatively easy to bound from above: The construction of *any* coupling between μ and ν yields a bound on the distance between μ and ν. In the same line of ideas, any C-Lipschitz mapping $f : \mathcal{X} \to \mathcal{X}'$ induces a C-Lipschitz mapping $P_p(\mathcal{X}) \to P_p(\mathcal{X}')$ (the proof is obvious).
5. Wasserstein distances incorporate a lot of the geometry of the space. For instance, the mapping $x \longmapsto \delta_x$ is an *isometry* between $\mathcal{X}$ and $P_p(\mathcal{X})$; but there are much deeper links. This partly explains why $P_p(\mathcal{X})$ is often very well adapted to statements that combine weak convergence and geometry.

To prove Theorem 6.7 I shall use the following lemma, which has interest on its own and will be useful again later.

Lemma 6.11 (Cauchy sequences in W_p are tight). *Let $\mathcal{X}$ be a Polish space, let $p \geq 1$ and let $(\mu_k)_{k\in\mathbb{N}}$ be a Cauchy sequence in $P_p(\mathcal{X})$, equipped with the distance W_p. Then (μ_k) is tight.*

The proof is not so obvious and the reader might skip it at first reading.

Proof of Lemma 6.11. Let $(\mu_k)_{k\in\mathbb{N}}$ be a Cauchy sequence in $P_p(\mathcal{X})$: This means that

$$W_p(\mu_k, \mu_\ell) \longrightarrow 0 \qquad \text{as } k, \ell \to \infty.$$

In particular,

$$\int d(x_0, x)^p \, d\mu_k(x) = W_p\big(\delta_{x_0}, \mu_k\big)^p \leq \Big[W_p(\delta_{x_0}, \mu_1) + W_p(\mu_1, \mu_k)\Big]^p$$

remains bounded as $k \to \infty$.

Since $W_p \geq W_1$, the sequence (μ_k) is also Cauchy in W_1 sense. Let $\varepsilon > 0$ be given, and let $N \in \mathbb{N}$ be such that

$$k \geq N \Longrightarrow W_1(\mu_N, \mu_k) < \varepsilon^2. \tag{6.9}$$

Then for any k, there exists $j \in \{1, \ldots, N\}$ such that $W_1(\mu_j, \mu_k) < \varepsilon^2$ (if $k \geq N$, this is (6.9); if $k < N$, just choose $j = k$).

Since the finite set $\{\mu_1, \ldots, \mu_N\}$ is tight, there is a compact set K such that $\mu_j[\mathcal{X} \setminus K] < \varepsilon$ for all $j \in \{1, \ldots, N\}$. By compactness, K can be covered by a finite number of small balls: $K \subset B(x_1, \varepsilon) \cup \ldots \cup B(x_m, \varepsilon)$.

Now write

$$U := B(x_1, \varepsilon) \bigcup \ldots \bigcup B(x_m, \varepsilon);$$

$$U_\varepsilon := \Big\{x \in \mathcal{X};\ d(x, U) < \varepsilon\Big\} \ \subset \ \ B(x_1, 2\varepsilon) \bigcup \ldots \bigcup B(x_m, 2\varepsilon);$$

$$\phi(x) := \left(1 - \frac{d(x, U)}{\varepsilon}\right)_+.$$

Note that $1_U \leq \phi \leq 1_{U_\varepsilon}$ and ϕ is $(1/\varepsilon)$-Lipschitz. By using these bounds and the Kantorovich–Rubinstein duality (6.3), we find that for $j \leq N$ and k arbitrary,

$$\begin{aligned}
\mu_k[U_\varepsilon] &\geq \int \phi \, d\mu_k \\
&= \int \phi \, d\mu_j + \left(\int \phi \, d\mu_k - \int \phi \, d\mu_j\right) \\
&\geq \int \phi \, d\mu_j - \frac{W_1(\mu_k, \mu_j)}{\varepsilon} \\
&\geq \mu_j[U] - \frac{W_1(\mu_k, \mu_j)}{\varepsilon}.
\end{aligned}$$

On one hand, $\mu_j[U]$ is at least $1 - \varepsilon$ if $j \leq N$; on the other hand, for each k we can find $j = j(k)$ such that $W_1(\mu_k, \mu_j) \leq \varepsilon^2$. So in fact

$$\mu_k[U_\varepsilon] \geq 1 - \varepsilon - \frac{\varepsilon^2}{\varepsilon} = 1 - 2\varepsilon.$$

At this point we have shown the following: For each $\varepsilon > 0$ there is a finite family $(x_i)_{1\leq i\leq m}$ such that all measures μ_k give mass at least $1-2\varepsilon$ to the set $Z := \cup B(x_i, 2\varepsilon)$. The

point is that Z might not be compact. There is a classical remedy: Repeat the reasoning with ε replaced by $\varepsilon 2^{-p+1}$, so there will be $(x_i)_{1\le i\le m(p)}$ such that

$$\mu_k\Big[\mathcal{X} \setminus \bigcup_{1\le i\le m(p)} B(x_i,\, \varepsilon 2^{-p})\Big] \le \varepsilon 2^{-p}.$$

It follows that

$$\mu_k[\mathcal{X} \setminus S] \le \varepsilon,$$

where

$$S := \bigcap_{1\le p\le\infty} \bigcup_{1\le i\le m(p)} \overline{B(x_i, \varepsilon 2^{-p})}.$$

By construction, S can be covered by finitely many balls of radius δ, where δ is arbitrarily small (just choose p large enough that $\varepsilon 2^{-p} < \delta$, and then $\overline{B(x_i, \varepsilon 2^{-p})}$ will be included in $B(x_i, \delta)$). Thus S is *totally bounded* (i.e. it can be covered by finitely many balls of arbitrarily small radius). It is also closed, as an intersection of finite unions of closed sets. Since $\mathcal{X}$ is a complete metric space, it follows from a classical result in topology that S is compact. This concludes the proof of Lemma 6.11. □

Proof of Theorem 6.7. Let $(\mu_k)_{k\in\mathbb{N}}$ be such that $\mu_k \to \mu$ in distance W_p; the goal is to show that μ_k converges to μ in $P_p(\mathcal{X})$. First, by Lemma 6.11, the sequence $(\mu_k)_{k\in\mathbb{N}}$ is tight, so there is a subsequence $(\mu_{k'})$ such that $\mu_{k'}$ converges weakly to some probability measure $\widetilde{\mu}$. Then by Lemma 4.2,

$$W_p(\widetilde{\mu}, \mu) \le \liminf_{k'\to\infty} W_1(\mu_{k'}, \mu) = 0.$$

So $\widetilde{\mu} = \mu$, and the whole sequence (μ_k) has to converge to μ. This however is not the end, since it only shows the weak convergence in the usual sense, not the convergence in $P_p(\mathcal{X})$.

For any $\varepsilon > 0$ there exists a constant $C_\varepsilon > 0$ such that for all nonnegative real numbers a, b,

$$(a+b)^p \le (1+\varepsilon)\, a^p + C_\varepsilon\, b^p.$$

Combining this inequality with the usual triangle inequality, we see that whenever x_0, x and y are three points in X, one has

$$d(x_0, x)^p \le (1+\varepsilon)\, d(x_0, y)^p + C_\varepsilon\, d(x, y)^p. \tag{6.10}$$

Now let (μ_k) be a sequence of probability measures in $P_p(\mathcal{X})$ such that $W_p(\mu_k, \mu) \longrightarrow 0$, and for each k, let π_k be an optimal transport plan between μ_k and μ. Integrating inequality (6.10) against π_k and using the marginal property, we find that

$$\int d(x_0, x)^p\, d\mu_k(x) \le (1+\varepsilon) \int d(x_0, y)^p\, d\mu(y) + C_\varepsilon \int d(x, y)^p\, d\pi_k(x, y).$$

But of course,

$$\int d(x, y)^p\, d\pi_k(x, y) = W_p(\mu_k, \mu)^p \xrightarrow[k\to\infty]{} 0.$$

Therefore,

$$\limsup_{k\to\infty} \int d(x_0, x)^p\, d\mu_k(x) \le (1+\varepsilon) \int d(x_0, x)^p\, d\mu(x).$$

Letting $\varepsilon \to 0$, we see that Property (ii) of Definition 6.6 holds true; so μ_k does converge weakly in $P_p(\mathcal{X})$ to μ.

Conversely, assume that μ_k converges weakly in $P_p(\mathcal{X})$ to μ; and again, for each k, introduce an optimal transport plan π_k between μ_k and μ. So

$$\int d(x,y)^p \, d\pi_k(x,y) \longrightarrow 0.$$

By Prokhorov's theorem, (μ_k) forms a tight sequence; also $\{\mu\}$ is tight. By Lemma 4.3, the sequence (π_k) is itself tight in $P(\mathcal{X} \times \mathcal{X})$. So, up to extraction of a subsequence, still denoted by (π_k), one may assume that

$$\pi_k \longrightarrow \pi \qquad \text{weakly in } P(\mathcal{X} \times \mathcal{X}).$$

Since each π_k is optimal, Theorem 5.18 guarantees that π is an optimal coupling of μ and μ, so this is the (completely trivial) coupling $\pi = (\mathrm{Id}\,, \mathrm{Id}\,)_\# \mu$ (in terms of random variables, $Y = X$). Since this is independent of the extracted subsequence, actually π is the limit of the whole sequence (π_k).

Now let $x_0 \in \mathcal{X}$ and $R > 0$. If $d(x,y) > R$, then at least one (say the largest) of the two numbers $d(x,x_0)$ and $d(x_0,y)$ has to be greater than $R/2$, and no less than $d(x,y)/2$. In a fancy writing,

$$1_{d(x,y) \geq R} \;\leq\; 1_{[d(x,x_0) \geq R/2 \text{ and } d(x,x_0) \geq d(x,y)/2]} \;+\; 1_{[d(x_0,y) \geq R/2 \text{ and } d(x_0,y) \geq d(x,y)/2]}.$$

So, obviously

$$\begin{aligned} \big[d(x,y)^p - R^p\big]_+ \;&\leq\; d(x,y)^p \, 1_{[d(x,x_0) \geq R/2 \text{ and } d(x,x_0) \geq d(x,y)/2]} \;+\; d(x,y)^p \, 1_{[d(x_0,y) \geq R/2 \text{ and } d(x_0,y) \geq d(x,y)/2]} \\ &\leq 2^p d(x,x_0)^p \, 1_{d(x,x_0) \geq R/2} \;+\; 2^p d(x_0,y)^p \, 1_{d(x_0,y) \geq R/2}. \end{aligned}$$

It follows that

$$\begin{aligned} W_p(\mu_k,\mu)^p &= \int d(x,y)^p \, d\pi_k(x,y) \\ &= \int \big[d(x,y) \wedge R\big]^p \, d\pi_k(x,y) + \int \big[d(x,y)^p - R^p\big]_+ \, d\pi_k(x,y) \\ &\leq \int \big[d(x,y) \wedge R\big]^p \, d\pi_k(x,y) + 2^p \int_{d(x,x_0) \geq R/2} d(x,x_0)^p \, d\pi_k(x,y) \\ &\qquad + 2^p \int_{d(x_0,y) > R/2} d(x_0,y)^p \, d\pi_k(x,y) \\ &= \int \big[d(x,y) \wedge R\big]^p \, d\pi_k(x,y) + 2^p \int_{d(x,x_0) \geq R/2} d(x,x_0)^p \, d\mu_k(x) \\ &\qquad + 2^p \int_{d(x_0,y) \geq R/2} d(x_0,y)^p \, d\nu_k(y). \end{aligned}$$

Since π_k converges weakly to π, the first term goes to 0 as $k \to \infty$. So

$$\limsup_{k \to \infty} W_p(\mu_k,\mu)^p \leq 2^p \limsup_{k \to \infty} \int_{d(x,x_0) \geq R/2} d(x,x_0)^p \, d\mu_k(x) + 2^p \limsup_{k \to \infty} \int_{d(x_0,y) \geq R/2} d(x_0,y)^p \, d\nu_k(y) = 0.$$

This concludes the argument. □

Control by total variation

The total variation is a classical notion of distance between probability measures. There is, by the way, a classical probabilistic representation formula for total variation,

$$\|\mu - \nu\|_{TV} \;=\; 2 \inf \mathbb{P}\,[X \neq Y], \tag{6.11}$$

where the infimum is over all couplings (X, Y) of (μ, ν), which can be seen as a very particular case of Kantorovich duality for the cost function $1_{x \neq y}$.

It seems natural that a control in Wasserstein distance should be weaker than a control in total variation. This is not completely true, because total variation does not take into account large distances. But one can control W_p by *weighted* total variation.

Theorem 6.12 (Wasserstein distances are controlled by weighted total variation). *Let μ and ν be two probability measures on a Polish space $(\mathcal{X}, d)$. Let $p \in [1, \infty)$ and $x_0 \in \mathcal{X}$. Then*

$$W_p(\mu, \nu) \leq 2^{\frac{1}{p'}} \left(\int d(x_0, x)^p \, d|\mu - \nu|(x) \right)^{\frac{1}{p}}, \qquad \frac{1}{p} + \frac{1}{p'} = 1. \tag{6.12}$$

Particular Case 6.13. In the case $p = 1$, if the diameter of $\mathcal{X}$ is bounded by D, this bound becomes $W_1(\mu, \nu) \leq D \, \|\mu - \nu\|_{TV}$.

Remark 6.14. The integral in the right-hand side of (6.12) can be interpreted as the Wasserstein distance W_1 for the particular cost function $[d(x_0, x) + d(x_0, y)] 1_{x \neq y}$.

Proof of Theorem 6.12. Let π be the transference plan obtained by keeping fixed all the mass shared by μ and ν, and distributing the rest uniformly: this is

$$\pi = (\mathrm{Id}\,, \mathrm{Id}\,)_\#(\mu \wedge \nu) + \frac{1}{a} \, (\mu - \nu)_+ \otimes (\mu - \nu)_-,$$

where $\mu \wedge \nu = \mu - (\mu - \nu)_+$ and $a = (\mu - \nu)_-[X] = (\mu - \nu)_+[X]$. A more sloppy but probably more readable way to write π is

$$\pi(dx\, dy) = (\mu \wedge \nu)(dx) \, \delta_{y=x} + \frac{1}{a} \, (\mu - \nu)_+(dx) \, (\mu - \nu)_-(dy).$$

By using the definition of W_p, the definition of π, the triangle inequality for d, the elementary inequality $(A + B)^p \leq 2^{p-1}(A^p + B^p)$, and the definition of a, we find that

$$\begin{aligned}
W_p(\mu, \nu)^p &\leq \int d(x, y)^p \, d\pi(x, y) \\
&= \frac{1}{a} \int d(x, y)^p \, d(\mu - \nu)_+(x) \, d(\mu - \nu)_-(y) \\
&\leq \frac{2^{p-1}}{a} \int \big[d(x, x_0)^p + d(x_0, y)^p \big] \, d(\mu - \nu)_+(x) \, d(\mu - \nu)_-(y) \\
&\leq 2^{p-1} \left[\int d(x, x_0)^p \, d(\mu - \nu)_+(x) + \int d(x_0, y)^p \, d(\mu - \nu)_-(y) \right] \\
&= 2^{p-1} \int d(x, x_0)^p \, d\big[(\mu - \nu)_+ + (\mu - \nu)_- \big](x) \\
&= 2^{p-1} \int d(x, x_0)^p \, d|\mu - \nu|(x).
\end{aligned}$$

□

Topological properties of the Wasserstein space

As a general rule, the Wasserstein space $P_p(\mathcal{X})$ inherits many properties of the basis space $\mathcal{X}$. Here is a first illustration:

Theorem 6.15 (Topological properties of the Wasserstein space). *Let $\mathcal{X}$ be a complete separable metric space and $p \in [1, +\infty)$. Then the Wasserstein space $P_p(\mathcal{X})$, metrized by the Wasserstein distance W_p, is also a complete separable metric space. In short: The Wasserstein space over a Polish space is itself a Polish space. Moreover, any probability measure can be approximated by a sequence of probability measures with finite support.*

Remark 6.16. If $\mathcal{X}$ is compact, then $P_p(\mathcal{X})$ is also compact; but if $\mathcal{X}$ is only locally compact, then $P_p(\mathcal{X})$ is *not* locally compact.

Proof of Theorem 6.15. The fact that $P_p(\mathcal{X})$ is a metric space was already explained, so let us turn to the proof of **separability**. Let $\mathcal{D}$ be a dense sequence in $\mathcal{X}$, and let $\mathcal{P}$ be the space of probability measures that can be written $\sum a_j \delta_{x_j}$, where the a_j are rational coefficients, and the x_j are finitely many elements in $\mathcal{D}$. It will turn out that $\mathcal{P}$ is dense in $P_p(\mathcal{X})$.

To prove this, let $\varepsilon > 0$ be given, and x_0 be an arbitrary element of $\mathcal{D}$. If μ lies in $P_p(\mathcal{X})$, then there exists a compact set $K \subset \mathcal{X}$ such that

$$\int_{\mathcal{X} \setminus K} d(x_0, x)^p \, d\mu(x) \le \varepsilon^p.$$

Cover K by a finite family of balls $B(x_k, \varepsilon/2)$, $1 \le k \le N$, with centers $x_k \in \mathcal{D}$. Then define

$$B'_k = B(x_k, \varepsilon) \setminus \bigcup_{j<k} B(x_j, \varepsilon).$$

Then all B'_k are disjoint and still cover K.

Define f on $\mathcal{X}$ by

$$f(B'_k \cap K) = \{x_k\}, \qquad f(\mathcal{X} \setminus K) = \{x_0\}.$$

Then, for all $x \in K$, $d(x, f(x)) \le \varepsilon$. So

$$\begin{aligned} \int d(x, f(x))^p \, d\mu(x) &\le \varepsilon^p \int_K d\mu(x) \; + \; \int_{\mathcal{X} \setminus K} d(x, x_0)^p \, d\mu(x) \\ &\le \varepsilon^p + \varepsilon^p = 2\, \varepsilon^p. \end{aligned}$$

Since $(\mathrm{Id}\,, f)$ is a coupling of μ and $f_\# \mu$, it follows that $W_p(\mu, f_\# \mu) \le 2\varepsilon^p$.

Of course, $f_\# \mu$ can be written as $\sum a_j \delta_{x_j}$, $0 \le j \le N$. This shows that μ might be approximated, with arbitrary precision, by a finite combination of Dirac masses. To conclude, it is sufficient to show that the coefficients a_j might be replaced by rational coefficients, up to a very small error in Wasserstein distance. So let D be a bound on the diameter of x_0, then by Theorem 6.12,

$$W_p \left(\sum_{j \le N} a_j \delta_{x_j}, \; \sum_{j \le N} b_j \delta_{x_j} \right) \; \le \; 2^{\frac{1}{p'}} D \sum_{j \le N} |a_j - b_j|^{\frac{1}{p}},$$

and obviously the latter quantity can be made as small as possible for some well-chosen rational coefficients b_j.

Now let us prove the **completeness**. Let again $(\mu_k)_{k\in\mathbb{N}}$ be a Cauchy sequence in $P_p(\mathcal{X})$. By Lemma 6.11, it admits a subsequence $(\mu_{k'})$ which converges weakly (in the usual sense) to some measure μ. Then,

$$\int d(x_0,x)^p\, d\mu(x) \le \liminf_{k\to\infty} \int d(x_0,x)^p\, d\mu_{k'}(x) < +\infty,$$

so μ belongs to $P_p(\mathcal{X})$. Moreover, by lower semi-continuity of W_p,

$$W_p(\mu,\mu_{\ell'}) \le \liminf_{k'\to\infty} W_p(\mu_{k'},\mu_{\ell'}),$$

so in particular

$$\limsup_{\ell'\to\infty} W_p(\mu,\mu_{\ell'}) \le \limsup_{k',\ell'\to\infty} W_p(\mu_{k'},\mu_{\ell'}) = 0,$$

which means that $\mu_{\ell'}$ converges to μ in W_p sense (and not just in the sense of weak convergence). So (μ_k) is a Cauchy sequence with a converging subsequence, and it follows by a classical argument that the whole sequence is converging. □

Bibliographical Notes

The terminology of Wasserstein distance is quite questionable, since (a) these distances were discovered and rediscovered by several authors throughout the twentieth century, including (in chronological order) Gini [186], Kantorovich [220], Wasserstein [361], Mallows [254] and Tanaka [349] (other important early contributors being Salvemini, Dall'Aglio, Hoeffding, Fréchet, Rubinstein, and maybe others); and (b) Wasserstein was only interested in the case $p = 1$. By the way, also the spelling of Wasserstein is doubtful: the original spelling was Vasershtein. All these issues are discussed in a historical note by Rüschendorf [316], who advocates the denomination of "minimal L^p-metric" instead of "Wasserstein distance".

In spite of these remarks, I will stick to the latter denomination, partly because it appears in many papers published since the end of the nineties (including my own), so that many researchers in partial differential equations or geometry agree on it. After all, even if this convention is a bit unfair since it does not give credit to all contributors, at least it does give credit to somebody.

Gini considered the special case where the random variables are discrete and lie on the real line; like Mallows later, he was interested by applications in statistics. Tanaka was interested by applications to partial differential equations. Both Mallows and Tanaka worked with the case $p = 2$, while Gini was interested both in $p = 1$ and $p = 2$, and Hoeffding and Fréchet worked with general p.

Wasserstein distances were used by Dobrushin in various areas of statistical mechanics, and Dobrushin advertised for them a lot, see e.g. [143, 144]. The Toscani metric is useful in the theory of the Boltzmann equation, see [364, Section 4.2] and references quoted therein. The Toscani metric and its variants are also rather handy for studying rates of convergence in the central limit theorem, or certain stable limit theorems [187]. The Lévy-Prokhorov metric appears in a number of textbooks, such as Dudley [147, p. 394]. For the taxonomy of probability metrics and their history, the unavoidable reference is the monograph by Rachev [307], which lists dozens and dozens of metrics together with

their main properties and applications. (Many of them are variants, particular cases or extensions of the Wasserstein and Lévy-Prokhorov metrics.)

Applications of the Wasserstein distances are too numerous to be quoted here; they will be encountered again in the sequel. For the moment let me just mention Tanaka's work on the Boltzmann equation [349, 350], which is reviewed in [365, Section 7.5]; and a surprising recent contribution by Werner [369], who showed that the Wasserstein distance is well adapted to quantify some variants of the uncertainty principle in quantum physics.

The equivalence between the four statements in Definition 6.6 is proven in [365, Theorem 7.12]. I borrowed the proof of Lemma 6.11 from Bolley [61]; and the scheme of proof of Theorem 6.7 from Ambrosio, Gigli and Savarè [15]. There are alternative proofs of Theorem 6.7 in the literature, for instance in [365].

The representation formula (6.11) for the total variation distance is a particular case of Strassen's duality theorem, see for instance [365, Section 1.4]. Remark 6.14 is taken from [188].

Theorem 6.12 is a copy-paste from [365, Proposition 7.10], which itself was a slight adaptation of [306, Lemma 10.2.3].

Theorem 6.15 belongs to folklore and has probably been proven many times; Sznitman [344] refers to Dobrushin for a proof of completeness. Other arguments are due to Rachev [307], and Ambrosio, Gigli and Savaré [15]. In the latter reference the proof is very simple but makes use of the deep Kolmogorov extension theorem. Here I followed a much more elementary argument due to Bolley [61].

The statement in Remark 6.16 is proven in [15, Remark 7.1.9].

7

Displacement interpolation

Now I shall discuss a **time-dependent** version of optimal transport leading to a *continuous* displacement of probability measures. There are several motivations for that extension:

- a time-dependent model gives a more complete description of the transport;
- the richer mathemtical structure will be useful later on.

As in the previous chapter I shall assume that the initial and final probability measures are defined on the same Polish space $(\mathcal{X}, d)$. The main additional structure assumption is that the cost is associated with an **action**, which is a way to measure the cost of displacement along a continuous curve, defined on a given time-interval, say $[0, 1]$. So the cost function between an initial point x and a final point y is obtained by minimizing the action among paths that go from x to y:

$$c(x, y) = \inf\Big\{\mathcal{A}(\gamma); \quad \gamma_0 = x, \ \gamma_1 = y; \quad \gamma \in \mathcal{C}\Big\}. \tag{7.1}$$

Here $\mathcal{C}$ is a certain class of continuous curves, to be specified in each particular case of interest, on which the action functional $\mathcal{A}$ is defined.

Of course, Assumption (7.1) is meaningless unless one requires some specific structure on the action functional (otherwise, just choose $\mathcal{A}(\gamma) = c(\gamma_0, \gamma_1)...$). A good notion of action should provide a recipe for choosing optimal paths, and in particular a recipe to interpolate between points in $\mathcal{X}$. It will turn out that under soft assumptions, this interpolation recipe between points can be "lifted" to an interpolation recipe between *probability measures*. This will provide a time-dependent notion of optimal transport, that will be called displacement interpolation (by opposition to the standard linear interpolation between probability measures).

This is a key chapter in these notes, and I have worked hard to attain a high level of generality, at the price of somewhat lengthy arguments. So the reader should not hesitate to skip proofs at first reading, concentrating on the statements and explanations.

Deterministic interpolation via action-minimizing curves

To better understand what an action functional should be, let us start with some examples and informal discussions. Consider a model where the unknown is the position of a given physical system in some position space, say a Riemannnian manifold M. (See the Appendix for reminders about Riemannian geometry if needed.) We learn from classical physics that in the absence of a potential, the action is the integral over time of the (instantaneous) kinetic energy:

$$\mathcal{A}(\gamma) = \int_0^1 \frac{|\dot\gamma_t|^2}{2}\, dt,$$

where $\dot\gamma_t$ stands for the velocity (or time-derivative) of the curve γ at time t. More generally, an action is classically given by the time-integral of a **Lagrangian** along the path:

$$\mathcal{A}(\gamma) = \int_0^1 L(\gamma_t, \dot\gamma_t, t)\, dt. \tag{7.2}$$

Here L is defined on $TM \times [0,1]$, where the smooth manifold M is the position space and the tangent bundle TM is the phase space, that is the space of all possible positions and velocities. It is natural to work in TM because one often deals with second-order differential equations on M (such as Newton's equations), which transform into first-order equations on TM. Typically L would take the form

$$L(x,v,t) = \frac{|v|^2}{2} - V(x) \tag{7.3}$$

where V is a potential; but much more complicated forms are admissible. When V is continuously differentiable, it is a simple particular case of the formula of first variation (recalled in the Appendix) that minimizers of (7.3), with given end-points, satisfy Newton's equation

$$\frac{d^2x}{dt^2} = -\nabla V(x). \tag{7.4}$$

To make sure that $\mathcal{A}(\gamma)$ is well-defined, it is natural to assume that the path γ is continuously differentiable, or piecewise continuously differentiable, or at least almost everywhere differentiable as a function of t. A classical and general setting is that of **absolutely continuous curves**: By definition, if $(\mathcal{X}, d)$ is a metric space, a continuous curve $\gamma : [0,1] \to \mathcal{X}$ is said to be absolutely continuous if there exists a function $\ell \in L^1([0,1]; dt)$ such that for all intermediate times $t_0 < t_1$ in $[0,1]$,

$$d(\gamma_{t_0}, \gamma_{t_1}) \le \int_{t_0}^{t_1} \ell(t)\, dt. \tag{7.5}$$

More generally, it is said to be absolutely continuous of order p if formula (7.5) holds with some $\ell \in L^p([0,1]; dt)$.

If γ is absolutely continuous, then the function $t \longmapsto d(\gamma_{t_0}, \gamma_t)$ is differentiable almost everywhere, and its derivative is integrable. But the converse is false: for instance, if γ is the "devil's staircase", encountered in measure theory textbooks (a nonconstant function whose distributional derivative is concentrated on the Cantor set in $[0,1]$), then γ is differentiable almost everywhere, and $\gamma'(t) = 0$ for almost every t, even though γ is not constant! This motivates the "integral" definition of absolute continuity based on formula (7.5).

If $\mathcal{X}$ is $\mathbb{R}^n$, or a smooth differentiable manifold, then absolutely continuous paths are differentiable for Lebesgue-almost all $t \in [0,1]$ (in physical words, the velocity is well-defined for almost all time).

Before going further, here are some simple and important examples. For all of them, the class of admissible curves is the space of absolutely continuous curves.

Example 7.1. In $\mathcal{X} = \mathbb{R}^n$, choose $L(x,v,t) = |v|$ (Euclidean norm of the velocity). Then the action is just the length functional, while the cost $c(x,y) = |x-y|$ is the Euclidean distance. Minimizing curves are straight lines, with arbitrary parametrization: $\gamma_t = \gamma_0 + s(t)(\gamma_1 - \gamma_0)$, where $s : [0,1] \to [0,1]$ is nondecreasing and absolutely continuous.

Example 7.2. In $\mathcal{X} = \mathbb{R}^n$ again, choose $L(x, v, t) = c(v)$, where c is strictly convex. By Jensen's inequality,

$$c(\gamma_1 - \gamma_0) = c\left(\int_0^1 \dot\gamma_t \, dt\right) \le \int_0^1 c(\dot\gamma_t)\, dt,$$

and this is an equality if and only if $\dot\gamma_t$ is constant. It follows that minimizers of the action are straight lines with *constant velocity*: $\gamma_t = \gamma_0 + t(\gamma_1 - \gamma_0)$. Then, of course,

$$c(x, y) = c(y - x).$$

Remark 7.3. This example shows that rather different Lagrangians can have the same minimizing curves.

Example 7.4. Let $\mathcal{X} = M$ be a smooth Riemannian manifold, TM its tangent bundle, and $L(x, v, t) = |v|^p$, $p \ge 1$. Then the cost function is $d(x, y)^p$, where d is the geodesic distance on M. There are two quite different cases:

- If $p > 1$, minimizing curves are defined by the equation $d^2\gamma_t/dt^2 = 0$ (zero acceleration), to be understood as $(d/dt)\dot\gamma_t = 0$, where (d/dt) stands for the covariant derivative along the path γ (once again, see the reminders in the Appendix if necessary). Such curves have constant speed ($(d/dt)|\dot\gamma_t| = 0$), and are called **minimizing, constant-speed geodesics**, or simply geodesics.

- If $p = 1$, minimizing curves are geodesic curves parametrized in an arbitrary way.

Example 7.5. Let again $\mathcal{X} = M$ be a smooth Riemannian manifold, and now consider a general Lagrangian $L(x, v, t)$, assumed to be *strictly convex* in the velocity variable v. The characterization and study of extremal curves for such Lagrangians, under various regularity assumptions, is one of the most classical topics in the calculus of variations. Here are some of the basic — which does not mean trivial — results in the field. In all the sequel, the Lagrangian L is a C^1 function defined on $TM \times [0, 1]$.

- By the first variation formula (whose proof is sketched in the Appendix), minimizing curves should satisfy the Euler-Lagrange equation

$$\frac{d}{dt}\Big[(\nabla_v L)(\gamma_t, \dot\gamma_t, t)\Big] = (\nabla_x L)(\gamma_t, \dot\gamma_t, t), \tag{7.6}$$

which is a generalization of (7.4). At least this equation should be satisfied for minimizing curves that are sufficiently smooth, say piecewise C^1.

- Assume that L is strictly convex and superlinear in the velocity variable, in the following sense:

$$\forall (x, t) \quad \begin{cases} v \longmapsto L(x, v, t) \quad \text{is convex} \\ \\ \dfrac{L(x, v, t)}{|v|} \xrightarrow[|v| \to \infty]{} +\infty \end{cases} \tag{7.7}$$

Then $v \longmapsto \nabla_v L$ is invertible, and (7.4) can be rewritten as a differential equation on the new unknown $\nabla_v L(\gamma, \dot\gamma, t)$.

- If there exists $K, C > 0$ such that

$$L(x, v, t) \ge K|v| - C,$$

then the action of a curve γ is bounded below by $KL(\gamma) - C$, where L is the length; it follows that all action-minimizing curves starting from a given compact K_0 and ending in a given compact K_1 stay within a bounded region.

- If minimizing curves depend smoothly on their position and velocity at some time, then there is also a bound on the velocities along minimizers that join K_0 to K_1. Indeed, there is a bound on $\int_0^1 L(x,v,t)\,dt$; so there is a bound on $L(x,v,t)$ for some t; so there is a bound on the velocity at some time, and then this bound is propagated in time.

- If in addition inequality $\nabla_v^2 L > 0$ holds true (more rigorously, $\nabla_v^2 L(x,\cdot,t) \geq K(x) g_x$ for all x, where g is the metric and $K(x) > 0$), then the new equation has locally Lipschitz coefficients, and the Cauchy-Lipschitz theorem can be applied to guarantee the unique local existence of Lipschitz continuous solutions to (7.6). Under the same assumptions on L, one can show directly that minimizers are of class at least C^1, and therefore satisfy (7.6). Conversely, solutions of (7.6) are locally (in time) minimizers of the action.

- Finally, the convexity of L makes it possible to introduce its Legendre transform (again, with respect to the velocity variable):

$$H(x,p,t) := \sup_{v \in T_x M} \Big(p \cdot v - L(x,v,t) \Big),$$

which is called the **Hamiltonian**; then one can recast (7.6) in terms of a Hamiltonian system, and access to the rich mathematical world of Hamiltonian dynamics. As soon as L is strictly convex superlinear, then the Legendre transform $(x,v) \to (x, \nabla_v L(x,v,t))$ is a homeomorphism, so assumptions about (x,v) can be re-expressed in terms of $(x, p = \nabla_v L(x,v,t))$.

- If L does not depend on t, then $H(x, \nabla_v L(x,v))$ is constant along minimizing curves $(x,v) = (\gamma_t, \dot\gamma_t)$; more generally, $(d/dt) H(x, \nabla_v L(x,v)) = (\partial_t H)(x, \nabla_v L(x,v))$.

Some of the above-mentioned assumptions will come back often in the sequel, so I shall summarize the most interesting ones in the following definition:

Definition 7.6 (Classical conditions on a Lagrangian function). *Let M be a Riemannian manifold, and $L(x,v,t)$ a Lagrangian on $TM \times [0,1]$. In this course, it is said that L satisfies the classical assumptions if*

(a) L is C^1 in all variables;

(b) At each (x,t), L is a strictly convex superlinear function of v, in the sense of (7.7)*;*

(c) There are constants $K, C > 0$ such that for all t, x, v, $L(x,v,t) \geq K|v| - C$;

(d) Minimizers are solutions of a well-defined locally Lipschitz flow; that is, there is a locally Lipschitz map $(x_0, v_0, t_0; t) \to X_t(x_0, v_0, t_0)$ on $TM \times [0,1] \times [0,1]$, such that each minimizer satisfies $\gamma(t) = X_t(t_0, \gamma(t_0), \dot\gamma(v_0))$.

The latter assumption is automatically satisfied if L is of class C^2, $\nabla_v^2 L > 0$ everywhere and L does not depend on t.

This looks general enough, however there are interesting cases where $\mathcal{X}$ does not have enough differentiable structure for the velocity vector to be well-defined (tangent spaces might not exist, for lack of smoothness). In such a case, it is still possible to define the *speed* along the curve:

$$|\dot\gamma_t| := \limsup_{\varepsilon \to 0} \frac{d(\gamma_t, \gamma_{t+\varepsilon})}{|\varepsilon|}. \tag{7.8}$$

This generalizes the natural notion of speed, which is the norm of the velocity vector. Thus it makes perfect sense to write a Lagrangian of the form $L(x, |v|, t)$ in a general metric space $\mathcal{X}$; here L might be essentially any measurable function on $\mathcal{X} \times \mathbb{R}_+ \times [0,1]$. (To ensure that $\int_0^1 L\,dt$ makes sense in $\mathbb{R} \cup \{+\infty\}$, it is natural to assume that L is bounded below.)

Example 7.7. Let $(\mathcal{X}, d)$ be a metric space. Define the **length** of an absolutely continuous curve by the formula

$$\mathcal{L}(\gamma) = \int_0^1 |\dot\gamma_t|\, dt. \tag{7.9}$$

Then minimizing curves are called geodesics. They may have variable speed, but, just as on a Riemannian manifold, one can always *reparametrize* them (that is, replace γ by $\widetilde\gamma$ where $\widetilde\gamma_t = \gamma_{s(t)}$, with s continuous increasing) in such a way that they have constant speed. In that case $d(\gamma_s, \gamma_t) = |t - s|\, \mathcal{L}(\gamma)$ for all $s, t \in [0, 1]$.

Example 7.8. Let again $(\mathcal{X}, d)$ be a metric space, but now consider the action

$$\mathcal{A}(\gamma) = \int_0^1 c(|\dot\gamma_t|)\, dt,$$

where c is strictly convex and strictly increasing (say $c(|v|) = |v|^p$, $p > 1$). Then,

$$c\big(d(\gamma_0, \gamma_1)\big) \le c(\mathcal{L}(\gamma)) = c\left(\int_0^1 |\dot\gamma_t|\, dt\right) \le \int_0^1 c(|\dot\gamma_t|)\, dt,$$

with equality in both inequalities if and only if γ is a constant-speed, minimizing geodesic. Thus $c(x, y) = c\big(d(x, y)\big)$ and minimizing curves are also geodesics, but with constant speed. Note that the distance can be recovered from the cost function, just by inverting c. As an illustration, if $p > 1$, and $c(|v|) = |v|^p$, then

$$d(x, y) = \inf\left\{\int_0^1 |\dot\gamma_t|^p\, dt; \quad \gamma_0 = x, \quad \gamma_1 = y\right\}^{\frac{1}{p}}.$$

In a given metric space, geodesics might not always exist, and it can even be the case that non-constant continuous curves do not exist (think of a discrete space). So to continue the discussion we shall have to impose appropriate assumptions on our metric space, and our cost function.

Here comes an important observation. When one wants to compute "in real life" the length of a curve, one does not use formula (7.9), but rather subdivides the curve into very small pieces, and approximates the length of each small piece by the distance between its endpoints. The finer the subdivision, the greater the measured approximate length (this is a consequence of the triangular inequality). So by taking finer and finer subdivisions we get an increasing family of measurements, whose upper bound may be taken as the measured length. This is actually an alternative definition of the length, which agrees with (7.9) for absolutely continuous curves, but does not require any further regularity assumption than plain continuity:

$$\mathcal{L}(\gamma) = \sup_{N \in \mathbb{N}} \sup_{0 = t_0 < t_1 < \dots < t_N = 1} \Big[d(\gamma_{t_0}, \gamma_{t_1}) + \cdots + d(\gamma_{t_{N-1}}, \gamma_{t_N})\Big]. \tag{7.10}$$

Then one can define a **length space** as a metric space $(\mathcal{X}, d)$ in which, for any two $x, y \in \mathcal{X}$,

$$d(x, y) = \inf_{\gamma \in C([0,1]; \mathcal{X})} \Big\{\mathcal{L}(\gamma); \quad \gamma_0 = x, \ \ \gamma_1 = y\Big\}. \tag{7.11}$$

If in addition $\mathcal{X}$ is complete and locally compact, then the infimum is a minimum, in which case the space is said to be strictly intrinsic. (By abuse of notation, one often says just "length space" for "strictly intrinsic length space"; another terminology is "geodesic space".) Such spaces play an important role in modern nonsmooth geometry.

Formulas (7.10) and (7.11) show an intimate link between the length and the distance: The length determines the distance by minimization, but conversely the distance determines the length by subdivision and approximation. The idea behind it is that the length of an "infinitesimal curve" is determined solely by the endpoints. A similar relation holds true for an action which is defined by a general Lagrangian of the form (7.2): indeed, if γ is differentiable at t, then

$$\int_t^{t+\varepsilon} L\big(\gamma_\tau, \dot\gamma_\tau, \tau\big)\, d\tau \simeq \varepsilon\, L(\gamma_t, \dot\gamma_t, t),$$

and the vector $\dot\gamma_t$ is uniquely determined, up to an error $O(\varepsilon)$, by γ_t and $\gamma_{t+\varepsilon}$. Such would not be the case if the function L would also depend on, say, the acceleration of the curve.

This **reconstruction property** plays an important role and it is natural to enforce it in an abstract generalization. To do so, it will be useful to consider an action as a *family* of functionals parametrized by the initial and the final times: So $\mathcal{A}^{s,t}$ is a functional on the set of paths $[s,t] \to \mathcal{X}$. Then we let

$$c^{s,t}(x,y) = \inf\Big\{\mathcal{A}^{s,t}(\gamma); \quad \gamma_s = x,\ \gamma_t = y; \quad \gamma^{s,t} \in \mathcal{C}\Big\}. \tag{7.12}$$

In words, $c^{s,t}(x,y)$ is the minimal work needed to go from point x at initial time s, to point y at final time t.

Example 7.9. Consider the Lagrangian $L(x,|v|,t) = |v|^p$. Then

$$c^{s,t}(x,y) = \frac{d(x,y)^p}{(t-s)^{p-1}}.$$

Note a characteristic property of these "power law" Lagrangians: The cost function depends on s,t only through multiplication by a constant. In particular, minimizing curves will be independent of s and t, up to reparametrization.

Abstract Lagrangian action

After all these preparations, the following definition should appear somewhat natural.

Definition 7.10 (Lagrangian action). *Let $(\mathcal{X}, d)$ be a Polish space, and let t_i, t_f be two real numbers. A Lagrangian action $(\mathcal{A})^{t_i,t_f}$ on $\mathcal{X}$ is a family of lower semi-continuous functionals $\mathcal{A}^{s,t}$ on $C([s,t];\mathcal{X})$ $(t_i \le s < t \le t_f)$, and cost functions $c^{s,t}$ on $\mathcal{X}\times\mathcal{X}$, such that*

(i) $0 \le t_1 < t_2 < t_3 \le 1 \Longrightarrow \quad \mathcal{A}^{t_1,t_2} + \mathcal{A}^{t_2,t_3} = \mathcal{A}^{t_1,t_3}$;

(ii) $\forall x,y \in \mathcal{X} \quad c^{s,t}(x,y) = \inf\Big\{\mathcal{A}^{s,t}(\gamma); \quad \gamma \in C([s,t];\mathcal{X});\ \gamma_s = x,\ \gamma_t = y\Big\}$;

(iii) For any curve $(\gamma_t)_{t_i\le t\le t_f}$,

$$\mathcal{A}^{t_i,t_f}(\gamma) = \sup_{N\in\mathbb{N}} \sup_{t_i=t_0\le t_1\le\ldots\le t_N=t_f} \Big[c^{t_0,t_1}(\gamma_{t_0},\gamma_{t_1}) + c^{t_1,t_2}(\gamma_{t_1},\gamma_{t_2}) + \cdots \\ \cdots + c^{t_{N-1},t_N}(\gamma_{t_{N-1}},\gamma_{t_N})\Big].$$

The functional $\mathcal{A} = \mathcal{A}^{t_i,t_f}$ will just be called the action, and the cost function $c = c^{t_i,t_f}$ the cost associated with the action. A curve $\gamma : [t_i,t_f] \to \mathcal{X}$ is said to be action-minimizing if it minimizes $\mathcal{A}$ among all curves having the same endpoints.

Example 7.11. (i) To recover (7.2) as a particular case of Definition 7.10, just set

$$\mathcal{A}^{s,t}(\gamma) = \int_s^t L(\gamma_\tau, \dot\gamma_\tau, \tau)\, d\tau. \tag{7.13}$$

(ii) A length space is a space in which $\mathcal{A}^{s,t}(\gamma) = \mathcal{L}(\gamma)$ (here $\mathcal{L}$ is the length) defines a Lagrangian action.

If $[t_i', t_f'] \subset [t_i, t_f]$ then it is clear that $(\mathcal{A})^{t_i,t_f}$ induces an action $(\mathcal{A})^{t_i',t_f'}$ on the time-interval $[t_i', t_f']$, just by restriction.

In the rest of this section I shall take $(t_i, t_f) = (0, 1)$, just for simplicity; of course one can always reduce to this case by reparametrization.

Now it will be useful to introduce further assumptions about existence and compactness of minimizing curves.

Definition 7.12 (coercive action). *Let $(\mathcal{A})^{0,1}$ be a Lagrangian action on a Polish space $\mathcal{X}$, with associated cost functions $(c^{s,t})_{0\le s<t\le 1}$. For any two times s, t $(0 \le s < t \le 1)$, and any two compacts $K_s, K_t \in \mathcal{X}$, let $\Gamma^{s,t}_{K_s\to K_t}$ be the set of minimizing paths that start in K_s at time s, and end in K_t at time t. The action will be called coercive if*

(i) It is bounded below, in the sense that

$$\inf_{s<t} \inf_{\gamma} \mathcal{A}^{s,t} > -\infty;$$

(ii) If $s < t$ are any two intermediate times, and K_s, K_t are any two nonempty compact sets such that $c^{s,t}(x, y) < +\infty$ for all $x \in K_s$, $y \in K_t$, then the set $\Gamma^{s,t}_{K_s\to K_t}$ is compact and nonempty.

In particular, minimizing curves between any two fixed points x, y with $c(x, y) < +\infty$ should always exist and form a compact set.

Remark 7.13. If each $\mathcal{A}^{s,t}$ has compact sub-level sets (that is, $\{\gamma;\ \mathcal{A}^{s,t}(\gamma) \le A\}$ is compact in $C([s, t]; \mathcal{X})$ for any $A \in \mathbb{R}$), then the lower semi-continuity of $\mathcal{A}^{s,t}$, together with a standard compactness argument (just as in Theorem 4.1) shows that there exists at least one action-minimizing curve among the set of curves that have prescribed final and initial points. In that case the requirement of nonemptiness in (ii) is fulfilled.

Example 7.14. (i) If $\mathcal{X}$ is a smooth Riemannian manifold and $L(x, v, t)$ is a Lagrangian satisfying the classical conditions of Definition 7.6, then the action defined by (7.13) is coercive.

(ii) If $\mathcal{X}$ is a geodesic length space, then the action defined by $\mathcal{A}^{s,t}(\gamma) = \mathcal{L}(\gamma)^2/(t - s)$ is coercive; in fact minimizers are constant-speed minimizing geodesic curves. On the other hand the action defined by $\mathcal{A}^{s,t}(\gamma) = \mathcal{L}(\gamma)$ is not coercive. Indeed, the possibility of reparametrization prevents the compactness of minimizing curves.

Proposition 7.15 (Properties of Lagrangian actions). *Let $(\mathcal{X}, d)$ be a Polish space and $(\mathcal{A})^{0,1}$ a coercive Lagrangian action on $\mathcal{X}$. Then*

(i) For all intermediate times $s \le t$, $c^{s,t}$ is lower semi-continuous on $\mathcal{X} \times \mathcal{X}$, with values in $\mathbb{R} \cup \{+\infty\}$;

(ii) If a curve γ on $[s, t] \subset [0, 1]$ is a minimizer of $\mathcal{A}^{s,t}$, then its restriction to $[s', t'] \subset [s, t]$ is also a minimizer for $\mathcal{A}^{s',t'}$;

(iii) For all times $t_1 < t_2 < t_3$ in $[0,1]$, and x_1, x_3 in $\mathcal{X}$,

$$c^{t_1,t_3}(x_1,x_3) = \inf_{x_2 \in \mathcal{X}} \Big(c^{t_1,t_2}(x_1,x_2) + c^{t_2,t_3}(x_2,x_3) \Big); \tag{7.14}$$

and if the infimum is achieved at some point x_2, then there is a minimizing curve which goes from x_1 at time t_1 to x_3 at time t_3, and passes through x_2 at time t_2.

(iv) A curve γ is a minimizer of $\mathcal{A}$ if and only if, for all intermediate times $t_1 < t_2 < t_3$ in $[0,1]$,

$$c^{t_1,t_3}(\gamma_{t_1},\gamma_{t_3}) = c^{t_1,t_2}(\gamma_{t_1},\gamma_{t_2}) + c^{t_2,t_3}(\gamma_{t_2},\gamma_{t_3}); \tag{7.15}$$

(v) If the cost functions $c^{s,t}$ are continuous, then the set Γ of all action-minimizing curves is closed in the uniform topology;

(vi) For all times $s < t$, there is a Borel map $S_{s\to t} : \mathcal{X} \times \mathcal{X} \to C([s,t];\mathcal{X})$, such that for all $x, y \in \mathcal{X}$, $S(x,y)$ belongs to $\Gamma^{s,t}_{x\to y}$. In words, there is a measurable recipe to join any two endpoints x and y by a minimizing curve $\gamma : [s,t] \to \mathcal{X}$.

Remark 7.16. The statement in (iv) is a powerful formulation of the minimizing property. It is often quite convenient from the technical point of view, even in a smooth setting, because it does not involve any time-derivative.

Remark 7.17. The continuity assumption in (v) is satisfied in most cases of interest. For instance, if $\mathcal{A}^{s,t}(\gamma) = \mathcal{L}(\gamma)^2/(t-s)$, then $c^{s,t}(x,y) = d(x,y)^2/(t-s)$, which is obviously continuous. Continuity also holds true in the other model example where $\mathcal{X}$ is a Riemannian manifold and the cost is obtained from a Lagrangian function $L(x,v,t)$ on $TM \times [0,1]$; a proof is sketched in the Appendix.

Proof of Proposition 7.15. Let us prove (i). By definition of the coercivity, $c(x,y)$ is never $-\infty$. Let $(x_k)_{k\in\mathbb{N}}$ and $(y_k)_{k\in\mathbb{N}}$ be sequences converging to x and y respectively. Then the family $(x_k) \cup \{x\}$ forms a compact set K_s, and the family $(y_k) \cup \{y\}$ also forms a compact set K_t. By assumption, for each k we can find a minimizing curve $\gamma_k : [s,t] \to \mathcal{X}$ joining x_k to y_k, so it lies within $\Gamma^{s,t}_{K_s \to K_t}$ which is compact. Thus from $(\gamma_k)_{k\in\mathbb{N}}$ we can extract a subsequence which converges uniformly to some minimizing curve γ. The uniform convergence implies that $x_k = \gamma_k(s) \to \gamma(s)$, $y_k = \gamma_k(t) \to \gamma(t)$, so γ joins x to y. The lower semi-continuity of $\mathcal{A}^{s,t}$ implies that $\mathcal{A}^{s,t}(\gamma) \le \liminf \mathcal{A}^{s,t}(\gamma_k)$, so

$$c^{s,t}(x,y) \le \mathcal{A}^{s,t}(\gamma) \le \liminf \mathcal{A}^{s,t}(\gamma_k) = \liminf c^{s,t}(x_k,y_k).$$

This establishes the lower semi-continuity of the cost $c^{s,t}$.

Property (ii) is obvious: if the restriction of γ to $[s',t']$ is not optimal, introduce $\widetilde{\gamma}$ on $[s',t']$ such that $\mathcal{A}^{s',t'}(\widetilde{\gamma}) < \mathcal{A}^{s',t'}(\gamma)$. Then the path obtained by concatenating γ on $[s,s']$, $\widetilde{\gamma}$ on $[s',t']$ and γ again on $[t',t]$, has a strictly lower action $\mathcal{A}^{s,t}$ than γ, which is impossible. (Obviously, this is the same line of reasoning as in the proof of the "restriction property" of Theorem 4.5.)

Now, to prove (iii), introduce minimizing curves $\gamma_{1\to 2}$ joining x_1 at time t_1, to x_2 at time t_2, and $\gamma_{2\to 3}$ joining x_2 at time t_2, to x_3 at time t_3. Then define γ on $[t_1,t_3]$ by concatenation of $\gamma_{1\to 2}$ and $\gamma_{2\to 3}$. From the axioms of Definition 7.10,

$$c^{t_1,t_3}(x_1,x_3) \le \mathcal{A}^{t_1,t_3}(\gamma) = \mathcal{A}^{t_1,t_2}(\gamma_{1\to 2}) + \mathcal{A}^{t_2,t_3}(\gamma_{2\to 3}) = c^{t_1,t_2}(x_1,x_2) + c^{t_2,t_3}(x_2,x_3).$$

The inequality in (ii) follows by taking the infimum over x_2. Moreover, if there is equality, that is,

$$c^{t_1,t_2}(x_1,x_2) + c^{t_2,t_3}(x_2,x_3) = c^{t_1,t_3}(x_1,x_3),$$

then equality holds everywhere in the above chain of inequalities, so the curve γ achieves the optimal cost $c^{t_1,t_3}(x_1,x_3)$, while passing through x_2 at time t_2.

It is a consequence of (ii) that any minimizer should satisfy (7.15), since the restrictions of γ to $[t_1,t_2]$ and to $[t_2,t_3]$ should both be minimizing. Conversely, let γ be a curve satisfying (7.15) for all (t_1,t_2,t_3) with $t_1 \le t_2 \le t_3$. By induction, it follows that for each subdivision $0 = t_0 \le t_1 \le \ldots \le t_N = 1$,

$$c^{0,1}(\gamma_0,\gamma_1) = \sum_j c^{t_j,t_{j+1}}(\gamma_{t_j},\gamma_{t_{j+1}}).$$

By point (iii) in Definition 7.10, it follows that $c^{0,1}(\gamma_0,\gamma_1) = \mathcal{A}^{0,1}(\gamma)$, which proves (iv).

If $0 \le t_1 < t_2 < t_3 \le 1$, let now $\Gamma(t_1,t_2,t_3)$ stand for the set of all curves satisfying (7.15). If all functions $c^{s,t}$ are continuous, then $\Gamma(t_1,t_2,t_3)$ is closed for the topology of uniform convergence. Then Γ is the intersection of all $\Gamma(t_1,t_2,t_3)$, so it is closed also; this proves statement (v). (Now there is a similarity with the proof of Theorem 5.18.)

For given times $s < t$, let $E_{s,t}$ be the "endpoints" mapping, defined on $C([s,t];\mathcal{X})$ by $\gamma \longmapsto (\gamma_s,\gamma_t)$. By assumption, any two points are joined by at least one minimizing curve, so $E_{s,t}$ is onto $\mathcal{X}\times\mathcal{X}$. It is clear that $E_{s,t}$ is a continuous map between Polish spaces, and by assumption $E_{s,t}^{-1}(x,y)$ is compact for all x,y. It follows by general theorems of "measurable selection" (see the references in the bibliographical notes) that $E_{s,t}$ admits a measurable right-inverse $S_{s\to t}$, i.e. $E_{s,t}\circ S_{s\to t} = \mathrm{Id}$. This proves statement (vi). □

Interpolation of random variables

Action-minimizing curves provide a fairly general framework to interpolate points, that can be seen as deterministic random variables. What happens when we want to interpolate between genuinely random variables, in a way that is most economic? Since a deterministic point can be identified with a Dirac mass, this new problem contains both the classical action-minimizing problem and the Monge–Kantorovich problem.

Here is a natural recipe. Let c be the cost associated with the Lagrangian action, and let μ_0, μ_1 be two given laws. First introduce an optimal coupling (X_0,X_1) of (μ_0,μ_1). Then introduce a random action-minimizing path $(X_t)_{0\le t\le 1}$ joining X_0 to X_1. (We shall see later that such a thing always exists.) Then the random variable X_t is an interpolation of X_0 and X_1; or equivalently the law μ_t is an interpolation of μ_0 and μ_1. This procedure is called **displacement interpolation**, by opposition to the linear interpolation $\mu_t = (1-t)\mu_0 + t\mu_1$. Note that there is a priori no uniqueness of the displacement interpolation.

One of the concepts that we just introduced deserves careful attention.

Definition 7.18 (dynamical optimal coupling). *Let $(\mathcal{X},d)$ be a metric space, $(\mathcal{A})$ a Lagrangian action on $\mathcal{X}$, c the associated cost, and Γ the set of action-minimizing curves. A dynamical optimal transference plan is a probability measure Π on Γ such that*

$$\pi_{0,1} := (e_0,e_1)_\# \Pi$$

is an optimal transference plan between μ_0 and μ_1. Equivalently, Π is the law of a random action-minimizing curve whose endpoints constitute an optimal coupling of μ_0 and μ_1.

A random curve whose law is a dynamical optimal transference plan between μ_0 and μ_1 is called a dynamical optimal coupling of (μ_0,μ_1).

The next theorem is the main result of this chapter. It shows that the law at time t of a dynamical optimal coupling can be seen as a minimizing path in the space of probability measures. In the important case when the cost is a power of the distance, the corollary stated right after the theorem shows that displacement interpolation can be thought of as a geodesic path in the space of probability measures. An informal slogan would be "A geodesic in the space of laws is the law of a geodesic"! The theorem also shows that such interpolations can be constructed under quite weak assumptions.

Theorem 7.19 (Displacement interpolation). *Let $(\mathcal{X}, d)$ be a Polish space, and $(\mathcal{A})^{0,1}$ a coercive Lagrangian action on $\mathcal{X}$, such that the associated cost functions $c^{s,t}$ are continuous. Whenever $0 \leq s < t \leq 1$, denote by $C^{s,t}(\mu, \nu)$ the optimal transport cost between the probability measures μ and ν with cost $c^{s,t}$; write $c = c^{0,1}$ and $C = C^{0,1}$. Let μ_0 and μ_1 be any two probability measures on $\mathcal{X}$, such that the optimal transport cost $C(\mu_0, \mu_1)$ is finite. Then, given a continuous path $(\mu_t)_{0 \leq t \leq 1}$, the following properties are equivalent:*

(i) For each $t \in [0, 1]$, μ_t is the law of γ_t, where $(\gamma_t)_{0 \leq t \leq 1}$ is a random action-minimizing curve such that (γ_0, γ_1) is an optimal coupling of (μ_0, μ_1);

(ii) For any three intermediate times $t_1 < t_2 < t_3$ in $[0, 1]$,

$$C^{t_1,t_2}(\mu_{t_1}, \mu_{t_2}) + C^{t_2,t_3}(\mu_{t_2}, \mu_{t_3}) = C^{t_1,t_3}(\mu_{t_1}, \mu_{t_3});$$

(iii) The path $(\mu_t)_{0 \leq t \leq 1}$ is a minimizing curve for the coercive action functional defined on $P(\mathcal{X})$ by

$$\mathbb{A}^{s,t}(\mu) = \sup_{N \in \mathbb{N}} \sup_{s = t_0 < t_1 < \ldots < t_N = t} \sum_{i=0}^{N-1} C^{t_i, t_{i+1}}(\mu_{t_i}, \mu_{t_{i+1}}) \tag{7.16}$$

$$= \inf_{\gamma} \mathbb{E}\, \mathcal{A}^{s,t}(\gamma), \tag{7.17}$$

where the last infimum is over all random curves $\gamma : [s, t] \to \mathcal{X}$ such that $\text{law}\,(\gamma_\tau) = \mu_\tau$ $(s \leq \tau \leq t)$.

In that case $(\mu_t)_{0 \leq t \leq 1}$ is said to be a displacement interpolation between μ_0 and μ_1. There always exists at least one such curve.

Finally, if $\mathcal{K}_0$ and $\mathcal{K}_1$ are two compact subsets of $P(\mathcal{X})$, such that $C^{0,1}(\mu_0, \mu_1) < +\infty$ for all $\mu_0 \in \mathcal{K}_0$, $\mu_1 \in \mathcal{K}_1$, then the set of dynamical optimal transference plans Π with $(e_0)_\# \Pi \in \mathcal{K}_0$ and $(e_1)_\# \Pi \in \mathcal{K}_1$ is compact.

Theorem 7.19 admits two important corollaries that are stated below:

Corollary 7.20 (displacement interpolation as geodesics). *Let $(\mathcal{X}, d)$ be a complete separable, locally compact length space. Let $p > 1$ and let $P_p(\mathcal{X})$ be the space of probability measures on $\mathcal{X}$ with finite moment of order p, metrized by the Wasserstein distance W_p. Then, given any two $\mu_0, \mu_1 \in P_p(\mathcal{X})$, and a continuous curve $(\mu_t)_{0 \leq t \leq 1}$, valued in $P(\mathcal{X})$, the following properties are equivalent:*

(i) μ_t is the law of γ_t, where γ is a random (minimizing, constant-speed) geodesic such that (γ_0, γ_1) is an optimal coupling;

(ii) $(\mu_t)_{0 \leq t \leq 1}$ is a geodesic curve in the space $P_p(\mathcal{X})$.

Moreover, if μ_0 and μ_1 are given, there always exists at least one such curve. More generally, if $\mathcal{K}_0 \subset P_p(\mathcal{X})$ and $\mathcal{K}_1 \subset P_p(\mathcal{X})$ are compact subsets of $P(\mathcal{X})$, then the set of geodesic curves $(\mu_t)_{0 \leq t \leq 1}$ with $\mu_0 \in \mathcal{K}_0$ and $\mu_1 \in \mathcal{K}_1$ is compact and nonempty; and also the set of dynamical optimal transference plans Π with $(e_0)_\# \Pi \in \mathcal{K}_0$, $(e_1)_\# \Pi \in \mathcal{K}_1$ is compact and nonempty.

Corollary 7.21 (Uniqueness of displacement interpolation). *With the same assumptions as in Theorem 7.19, if*

(a) there is a unique optimal transference plan π between μ_0 and μ_1;

(b) $\pi(dx_0\, dx_1)$-almost surely, there is a unique action-minimizing curve joining x_0 to x_1;

then there is a unique displacement interpolation $(\mu_t)_{0\le t\le 1}$ joining μ_0 to μ_1.

Remark 7.22. In Corollary 7.20, $\mathcal{A}^{s,t}(\gamma) = \int_s^t |\dot\gamma_\tau|^p\, d\tau$. Then action-minimizing curves in $\mathcal{X}$ are the same, whatever the value of $p > 1$. Yet geodesics in $P_p(\mathcal{X})$ are not the same for different values of p, because a coupling of (μ_0, μ_1) which is optimal for a certain value of p, might well not be for another value.

Remark 7.23. Theorem 7.19 applies just as well to Lagrangian functions $L(x, v, t)$ on a Riemannian manifold TM, as soon as L is C^2 and satisfies the classical conditions of Definition 7.6. Then μ_t is the law at time t of a random solution of the Euler-Lagrange equation (7.6).

Remark 7.24. In Theorem 7.19, the minimizing property of the path (μ_t) is expressed in a weak formulation, which makes sense with a lot of generality. This theorem however leaves open certain natural questions:

- Is there a differential equation for geodesic curves, or more generally optimal paths $(\mu_t)_{0\le t\le 1}$? Of course, the answer to that question is related to the possibility of defining a tangent space in the space of measures.

- Is there a more explicit formula for the action on the space of probability measures, say for a simple enough action on $\mathcal{X}$? Can it be written as $\int_0^1 \mathbb{L}(\mu_t, \dot\mu_t, t)\, dt$? Of course, in Corollary 7.20 this is the case with $\mathbb{L} = |\dot\mu|^p$, but this expression is not very "explicit".

- Are geodesic paths non-branching? (Does the velocity at initial time uniquely determine the final measure μ_1?)

- Can one identify simple conditions for the existence of a *unique* geodesic path between two given probability measures?

All these questions will be answered affirmatively in the sequel of these notes, under suitable regularity assumptions on the space, the action or the probability measures.

Remark 7.25. The assumption of local compactness in Corollary 7.20 is not superficial: it is used to guarantee the coercivity of the action. For spaces that are not locally compact, there might be an analogous theory, but it is certainly more tricky. First of all, selection theorems are not immediately available if one does not assume compactness of the set of geodesics joining two given endpoints. More importantly, the convergence scheme used below to construct a random geodesic curve from a time-dependent law might fail to work. Here we are encountering a general principle in probability theory: Analytic characterizations of stochastic processes (like those based on semigroups, generators, etc.) are essentially available only in locally compact spaces. In spite of all that, there are some representation theorems for Wasserstein geodesics that do not need local compactness; see the bibliographical notes for details.

The proof of Theorem 7.19 is not so difficult, but a bit cumbersome because of measurability issues. For training purposes, the reader might rewrite it in the simpler case where any pair of points is joined by a *unique* geodesic (as in the case of $\mathbb{R}^n$). To help understanding, I shall first sketch the main idea.

Main idea in the proof of Theorem 7.19. The delicate part consists in showing that if (μ_t) is a given action-minimizing curve, then there exists a random minimizer γ such that $\mu_t = \text{law}\,(\gamma_t)$. This γ will be constructed by dyadic approximation, as follows. First let $(\gamma_0^{(0)}, \gamma_1^{(0)})$ be an optimal coupling of (μ_0, μ_1). (Here the notation $\gamma_0^{(0)}$ could be replaced by just x_0, it does not mean that there is some curve $\gamma^{(0)}$ behind.) Then let $(\gamma_0^{(1)}, \gamma_{1/2}^{(1)})$ be an optimal coupling of $(\mu_0, \mu_{1/2})$, and $((\gamma')_{1/2}^{(1)}, \gamma_1^{(1)})$ be an optimal coupling of $(\mu_{1/2}, \mu_1)$. By gluing these couplings together, I can actually assume that $(\gamma')_{1/2}^{(1)} = \gamma_{1/2}^{(1)}$, so that I have a triple $(\gamma_0^{(1)}, \gamma_{1/2}^{(1)}, \gamma_1^{(1)})$ in which the first two components on one hand, and the last two components on the other hand, constitute optimal couplings.

Now the key observation is that if $(\gamma_{t_1}, \gamma_{t_2})$ and $(\gamma_{t_2}, \gamma_{t_3})$ are optimal couplings of (μ_{t_1}, μ_{t_2}) and (μ_{t_2}, μ_{t_3}) respectively, and the μ_{t_k} satisfy the equality appearing in (ii), then also $(\gamma_{t_1}, \gamma_{t_3})$ should be optimal. Indeed, by taking expectation in the inequality

$$c^{t_1,t_3}(\gamma_{t_1}, \gamma_{t_3}) \le c^{t_1,t_2}(\gamma_{t_1}, \gamma_{t_2}) + c^{t_2,t_3}(\gamma_{t_2}, \gamma_{t_3})$$

and using the optimality assumption, one obtains

$$\mathbb{E}\, c^{t_1,t_3}(\gamma_{t_1}, \gamma_{t_3}) \le C^{t_1,t_2}(\mu_{t_1}, \mu_{t_2}) + C^{t_2,t_3}(\mu_{t_2}, \mu_{t_3}).$$

Now the fact that (μ_t) is action-minimizing imposes

$$C^{t_1,t_2}(\mu_{t_1}, \mu_{t_2}) + C^{t_2,t_3}(\mu_{t_2}, \mu_{t_3}) = C^{t_1,t_3}(\mu_{t_1}, \mu_{t_3}).$$

So actually

$$\mathbb{E}\, c^{t_1,t_3}(\gamma_{t_1}, \gamma_{t_3}) \le C^{t_1,t_3}(\mu_{t_1}, \mu_{t_3}),$$

which means that indeed $(\gamma_{t_1}, \gamma_{t_3})$ is an optimal coupling of (μ_{t_1}, μ_{t_3}) for the cost function c^{t_1,t_3}.

So $(\gamma_0^{(1)}, \gamma_1^{(1)})$ is an optimal coupling of (μ_0, μ_1). Now we can proceed in the same manner and define, for each k, a random discrete path $(\gamma_{j\,2^{-k}}^{(k)})$ such that $(\gamma_s^{(k)}, \gamma_t^{(k)})$ is an optimal coupling for all times s, t of the form $j/2^k$. These are only discrete paths, but it is possible to extend them into paths $(\gamma_t^{(k)})_{0 \le t \le 1}$ that are minimizers of the action. Of course, if t is not of the form $j/2^k$, there is no reason why law $(\gamma_t^{(k)})$ would coincide with μ_t. But now we shall be able to pass to the limit as $k \to \infty$, for each dyadic time, and then conclude by a density argument. □

Complete proof of Theorem 7.19. First, if $\mathcal{A}^{s,t}(\gamma)$ is bounded below by a constant $-C$, independently of s, t and γ, then the same is true of the cost functions $c^{s,t}$ and of the total costs $C^{s,t}$. So all the quantities appearing in the proof will be well-defined, the value $+\infty$ being possibly attained. Moreover, the action $\mathbb{A}^{s,t}$ defined by the formula in (iii) will also be bounded below by the same constant $-C$, so Property (i) of Definition 7.12 will be satisfied.

Let now μ_0 and μ_1 be given. According to Theorem 4.1, there exists at least one optimal transference plan π between μ_0 and μ_1, for the cost $c = c^{0,1}$. Let $S_{0\to 1}$ be the mapping appearing in Proposition 7.15(vi), and define

$$\Pi := (S_{0\to 1})_{\#}\pi.$$

Then Π defines the law of a random geodesic γ, and the identity $E_{0,1} \circ S_{0\to 1} = \text{Id}$ implies that the end-points of γ are distributed according to π. This proves the existence of a

path satisfying (i). Now the main part of the proof consists in checking the equivalence of properties (i) and (ii). This will be performed in four steps.

Step 1. Let $(\mu_t)_{0\le t\le 1}$ be any continuous curve in the space of probability measures, and let t_1, t_2, t_3 be three intermediate times. Let $\pi_{t_1\to t_2}$ be an optimal transference plan between μ_{t_1} and μ_{t_2} for the transport cost c^{t_1,t_2}, and similarly let $\pi_{t_2\to t_3}$ be an optimal transference plan between μ_{t_2} and μ_{t_3} for the transport cost c^{t_2,t_3}. By the Gluing Lemma of Chapter 1 one can construct random variables $(\gamma_{t_1}, \gamma_{t_2}, \gamma_{t_3})$ such that law $(\gamma_{t_1}, \gamma_{t_2}) = \pi_{t_1\to t_2}$ and law $(\gamma_{t_2}, \gamma_{t_3}) = \pi_{t_2\to t_3}$ (in particular, law $(\gamma_{t_i}) = \mu_{t_i}$ for $i = 1, 2, 3$). Then, by (7.14),

$$\begin{aligned} C^{t_1,t_3}(\mu_{t_1}, \mu_{t_3}) &\le \mathbb{E}\, c^{t_1,t_3}(\gamma_{t_1}, \gamma_{t_3}) \le \mathbb{E}\, c^{t_1,t_2}(\gamma_{t_1}, \gamma_{t_2}) + \mathbb{E}\, c^{t_2,t_3}(\gamma_{t_2}, \gamma_{t_3}) \\ &= C^{t_1,t_2}(\mu_{t_1}, \mu_{t_2}) + C^{t_2,t_3}(\mu_{t_2}, \mu_{t_3}). \end{aligned}$$

This inequality holds for any path, optimal or not.

Step 2. Assume that (μ_t) satisfies (i), so there is a dynamical optimal transference plan Π such that $\mu_t = (e_t)_\# \Pi$. Let γ be a random minimizing curve with law Π, and consider the obvious coupling $(\gamma_{t_1}, \gamma_{t_2})$ (resp. $(\gamma_{t_2}, \gamma_{t_3})$) of (μ_{t_1}, μ_{t_2}) (resp. (μ_{t_2}, μ_{t_3})). Then from the definition of the optimal cost and the minimizing property of γ,

$$\begin{aligned} C^{t_1,t_2}(\mu_{t_1}, \mu_{t_2}) + C^{t_2,t_3}(\mu_{t_2}, \mu_{t_3}) &\le \mathbb{E}\, c^{t_1,t_2}(\gamma_{t_1}, \gamma_{t_2}) + \mathbb{E}\, c^{t_2,t_3}(\gamma_{t_2}, \gamma_{t_3}) \\ &= \mathbb{E}\, \mathcal{A}^{t_1,t_2}(\gamma) + \mathbb{E}\, \mathcal{A}^{t_2,t_3}(\gamma) = \mathbb{E}\, \mathcal{A}^{t_1,t_3}(\gamma) = \mathbb{E}\, c^{t_1,t_3}(\gamma_{t_1}, \gamma_{t_3}). \quad (7.18) \end{aligned}$$

Now choose $t_1 = 0$, $t_2 = t$, $t_3 = 1$. Since by assumption (γ_0, γ_1) is an optimal coupling of (μ_0, μ_1), the above computation implies

$$C^{0,t}(\mu_0, \mu_t) + C^{t,1}(\mu_t, \mu_1) \le C^{0,1}(\mu_0, \mu_1),$$

and since the reverse inequality holds as a consequence of Step 1, actually

$$C^{0,t}(\mu_0, \mu_t) + C^{t,1}(\mu_t, \mu_1) = C^{0,1}(\mu_0, \mu_1).$$

Moreover, equality has to hold in (7.18) (for that particular choice of intermediate times), so that $C^{0,t}(\mu_0, \mu_t) = \mathbb{E}\, c^{0,t}(\gamma_0, \gamma_t)$, which means that (γ_0, γ_t) should actually be an optimal coupling of (μ_0, μ_t); similarly, (γ_t, γ_1) should be an optimal coupling of (μ_t, μ_1).

Next choose $t_1 = 0$, $t_2 = s$, $t_3 = t$, and apply the previous deduction to discover that (γ_s, γ_t) is an optimal coupling of (μ_s, μ_t). After inserting this information in (7.18) with $s = t_2$ and $t = t_3$, we recover

$$C^{t_1,t_2}(\mu_{t_1}, \mu_{t_2}) + C^{t_2,t_3}(\mu_{t_2}, \mu_{t_3}) \le C^{t_1,t_3}(\mu_{t_1}, \mu_{t_3}).$$

This together with Step 1 proves that (μ_t) satisfies Property (ii). So far we have proven (i) $\Rightarrow$ (ii).

Step 3. Assume that (μ_t) satisfies Property (ii); then we can perform again the same computation as in Step 1, but now all the inequalities have to be equalities. This implies that the random variables $(\gamma_{t_1}, \gamma_{t_2}, \gamma_{t_3})$ satisfy:

(a) $(\gamma_{t_1}, \gamma_{t_3})$ is an optimal coupling of (μ_{t_1}, μ_{t_3}) for the cost c^{t_1,t_3};

(b) $c^{t_1,t_3}(\gamma_{t_1}, \gamma_{t_3}) = c^{t_1,t_2}(\gamma_{t_1}, \gamma_{t_2}) + c^{t_2,t_3}(\gamma_{t_2}, \gamma_{t_3})$ almost surely.

Armed with that information, we proceed as follows. We start from an optimal coupling (γ_0, γ_1) of (μ_0, μ_1), with joint law $\pi_{0\to 1}$. Then as in Step 1 we construct a triple $(\gamma_0^{(1)}, \gamma_{\frac12}^{(1)}, \gamma_1^{(1)})$ with law $(\gamma_0^{(1)}) = \mu_0$, law $(\gamma_{\frac12}^{(1)}) = \mu_{\frac12}$, law $(\gamma_1^{(1)}) = \mu_1$, such that $(\gamma_0^{(1)}, \gamma_{\frac12}^{(1)})$

is an optimal coupling of $(\mu_0, \mu_{\frac12})$ for the cost $c^{0,\frac12}$ and $(\gamma^{(1)}_{\frac12}, \gamma^{(1)}_1)$ is an optimal coupling of $(\mu_{\frac12}, \mu_1)$ for the cost $c^{\frac12,1}$. From (a) and (b) above we know that $(\gamma^{(1)}_0, \gamma^{(1)}_1)$ is an optimal coupling of (μ_0, μ_1) (but law $(\gamma^{(1)}_0, \gamma^{(1)}_1)$ might be different from law (γ_0, γ_1)), and moreover $c^{0,1}(\gamma^{(1)}_0, \gamma^{(1)}_1) = c^{0,\frac12}(\gamma^{(1)}_0, \gamma^{(1)}_{\frac12}) + c^{\frac12,1}(\gamma^{(1)}_{\frac12}, \gamma^{(1)}_1)$ almost surely.

Now it is possible to iterate the construction, introducing more and more midpoints. By a reasoning similar to the one above and an induction argument, one can construct, for each integer $k \geq 1$, random variables $(\gamma^{(k)}_0, \gamma^{(k)}_{\frac{1}{2^k}}, \gamma^{(k)}_{\frac{2}{2^k}}, \gamma^{(k)}_{\frac{3}{2^k}}, \dots \gamma^{(k)}_1)$ in such a way that

(a) for any two $i, j \leq 2^k$, $(\gamma^{(k)}_{\frac{i}{2^k}}, \gamma^{(k)}_{\frac{j}{2^k}})$ constitutes an optimal coupling of $(\mu_{\frac{i}{2^k}}, \mu_{\frac{j}{2^k}})$,

(b) for any three indices $i_1, i_2, i_3 \leq 2^k$, one has

$$c^{\frac{i_1}{2^k}, \frac{i_3}{2^k}}\big(\gamma^{(k)}_{\frac{i_1}{2^k}}, \gamma^{(k)}_{\frac{i_3}{2^k}}\big) = c^{\frac{i_1}{2^k}, \frac{i_2}{2^k}}\big(\gamma^{(k)}_{\frac{i_1}{2^k}}, \gamma^{(k)}_{\frac{i_2}{2^k}}\big) + c^{\frac{i_2}{2^k}, \frac{i_3}{2^k}}\big(\gamma^{(k)}_{\frac{i_2}{2^k}}, \gamma^{(k)}_{\frac{i_3}{2^k}}\big).$$

At this stage it is convenient to extend the random variables $\gamma^{(k)}$, which are only defined for times $j/2^k$, into (random) continuous curves $(\gamma^{(k)}_t)_{0\leq t\leq 1}$. For that we use Proposition 7.15(vi) again, and for $t \in (i/2^k, (i+1)/2^k)$ we define

$$\gamma_t := e_t\Big(S_{\frac{i}{2^k}, \frac{i+1}{2^k}}(\gamma_{\frac{i}{2^k}}, \gamma_{\frac{i+1}{2^k}})\Big).$$

(Recall that e_t is just the evaluation at time t.) Then the law $\Pi^{(k)}$ of $(\gamma_t)_{0\leq t\leq 1}$ is a probability measure on the set of continuous curves in $\mathcal{X}$.

Now I claim that $\Pi^{(k)}$ is actually concentrated on minimizing curves (Skip at first reading and go directly to Step 4.) To prove this, it is sufficient to check the criterion in Proposition 7.15(iv), involving three intermediate times t_1, t_2, t_3. By construction, the criterion holds true if all these times belong to the same time-interval $[i/2^k, (i+1)/2^k]$, and also if they are all of the form $j/2^k$; the problem consists in "crossing subintervals". Let us show that

$$\begin{aligned}
&(i-1)\,2^{-k} < s < i\,2^{-k} \leq j\,2^{-k} < t < (j+1)\,2^{-k} \\
\Longrightarrow\quad &\begin{cases} c^{\frac{i-1}{2^k}, \frac{i+1}{2^k}}\big(\gamma_{\frac{i-1}{2^k}}, \gamma_{\frac{i+1}{2^k}}\big) = c^{\frac{i-1}{2^k}, s}\big(\gamma_{\frac{i-1}{2^k}}, \gamma_s\big) + c^{s,t}(\gamma_s, \gamma_t) + c^{t, \frac{j+1}{2^k}}\big(\gamma_t, \gamma_{\frac{i+1}{2^k}}\big) \\ \\ c^{s,t}(\gamma_s, \gamma_t) = c^{s, \frac{i}{2^k}}(\gamma_s, \gamma_{\frac{i}{2^k}}) + c^{\frac{i}{2^k}, \frac{j}{2^k}}(\gamma_{\frac{i}{2^k}}, \gamma_{\frac{j}{2^k}}) + c^{\frac{j}{2^k}, t}(\gamma_{\frac{j}{2^k}}, \gamma_t) \end{cases}
\end{aligned} \tag{7.19}$$

To prove this, we start with

$$\begin{aligned}
c^{\frac{i-1}{2^k}, \frac{i+1}{2^k}}(\gamma_{\frac{i-1}{2^k}}, \gamma_{\frac{i+1}{2^k}}) &\leq c^{\frac{i-1}{2^k}, s}(\gamma_{\frac{i-1}{2^k}}, \gamma_s) + c^{s,t}(\gamma_s, \gamma_t) + c^{t, \frac{i+1}{2^k}}(\gamma_t, \gamma_{\frac{i+1}{2^k}}) \\
\leq c^{\frac{i-1}{2^k}, s}(\gamma_{\frac{i-1}{2^k}}, \gamma_s) \;+\; c^{s, \frac{i}{2^k}}(\gamma_s, \gamma_{\frac{i}{2^k}}) \;+\; &c^{\frac{i}{2^k}, \frac{i+1}{2^k}}(\gamma_{\frac{i}{2^k}}, \gamma_{\frac{i+1}{2^k}}) \\
&+ \;\dots\; + \; c^{\frac{j}{2^k}, t}(\gamma_{\frac{j}{2^k}}, \gamma_t) + c^{t, \frac{j+1}{2^k}}(\gamma_t, \gamma_{\frac{j+1}{2^k}}).
\end{aligned} \tag{7.20}$$

Since we have used minimizing curves to interpolate on each dyadic subinterval,

$$c^{\frac{i-1}{2^k}, s}(\gamma_{\frac{i-1}{2^k}}, \gamma_s) + c^{s, \frac{i}{2^k}}(\gamma_s, \gamma_{\frac{i}{2^k}}) = c^{\frac{i-1}{2^k}, \frac{i}{2^k}}(\gamma_{\frac{i-1}{2^k}}, \gamma_{\frac{i}{2^k}}),$$

etc. So the right-hand side of (7.20) coincides with

$$c^{\frac{i-1}{2^k},\frac{i}{2^k}}(\gamma_{\frac{i-1}{2^k}},\gamma_{\frac{i}{2^k}}) + \ldots + c^{\frac{j}{2^k},\frac{j+1}{2^k}}(\gamma_{\frac{j}{2^k}},\gamma_{\frac{j+1}{2^k}}),$$

and we know from the construction of $\mathit{\Pi}^{(k)}$ that this is just $c^{\frac{i-1}{2^k},\frac{j+1}{2^k}}(\gamma_{\frac{i-1}{2^k}},\gamma_{\frac{j+1}{2^k}})$. So there has to be equality everywhere in (7.20), which leads to (7.19). After that it is an easy game to conclude the proof of the minimizing property for arbitrary times t_1, t_2, t_3.

Step 4. To recapitulate: Starting from a curve $(\mu_t)_{0\le t\le 1}$, we have constructed a family of probability measures $\mathit{\Pi}^{(k)}$ which are all concentrated on the set $\mathit{\Gamma}$ of minimizing curves, and satisfy $(e_t)_\#\mathit{\Pi}^{(k)} = \mu_t$ for all $t = j/2^k$. It remains to pass to the limit as $k \to \infty$. For that we shall check the tightness of the sequence $(\mathit{\Pi}^{(k)})_{k\in\mathbb{N}}$. Let $\varepsilon > 0$ be arbitrary. Since μ_0, μ_1 are tight, there are compact sets K_0, K_1 such that $\mu_0[\mathcal{X}\setminus K_0] \le \varepsilon$, $\mu_1[\mathcal{X}\setminus K_1] \le \varepsilon$. From the coercivity of the action, the set $\mathit{\Gamma}^{0,1}_{K_0\to K_1}$ of action-minimizing curves joining K_0 to K_1 is compact, and $\mathit{\Pi}\big[\mathit{\Gamma}\setminus \mathit{\Gamma}^{0,1}_{K_0\to K_1}\big]$ is (with obvious notation)

$$\mathbb{P}\left[(\gamma_0,\gamma_1)\notin K_0\times K_1\right] \le \mathbb{P}\left[\gamma_0\notin K_0\right] + \mathbb{P}\left[\gamma_1\notin K_1\right] = \mu_0[\mathcal{X}\setminus K_0] + \mu_1[\mathcal{X}\setminus K_1] \le 2\varepsilon.$$

This proves the tightness of the family $(\mathit{\Pi}^{(k)})$. So one can extract a subsequence thereof, still denoted $\mathit{\Pi}^{(k)}$, that converges weakly to some probability measure $\mathit{\Pi}$.

By Proposition 7.15(v), $\mathit{\Gamma}$ is closed; so $\mathit{\Pi}$ is still supported in $\mathit{\Gamma}$. Moreover, for all dyadic time $t = i/2^\ell$ in $[0,1]$, we have, if k is larger than ℓ, $(e_t)_\#\mathit{\Pi}^{(k)} = \mu_t$, and by passing to the limit we find that $(e_t)_\#\mathit{\Pi} = \mu_t$ also.

By assumption, μ_t depends continuously on t. So, to conclude that $(e_t)_\#\mathit{\Pi} = \mu_t$ for all times $t \in [0,1]$ it now suffices to check the continuity of $(e_t)_\#\mathit{\Pi}$ as a function of t. In other words, if φ is an arbitrary bounded continuous function on $\mathcal{X}$, one has to show that

$$\psi(t) = \mathbb{E}\,\varphi(\gamma_t)$$

is a continuous function of t if γ is a random geodesic with law $\mathit{\Pi}$. But this is a simple consequence of the continuity of $t \longmapsto \gamma_t$ (for all γ), and Lebesgue's dominated convergence theorem. This concludes Step 4, and the proof of (ii) $\Rightarrow$ (i).

Next, let us check that the two expressions for $\mathcal{A}^{s,t}$ in (iii) do coincide. This is about the same computation as in Step 1 above. Let $s < t$ be given, let $(\mu_\tau)_{s\le\tau\le t}$ be a continuous path, and let (t_i) be a subdivision of $[s,t]$. Let further γ be such that law $(\gamma_\tau) = \mu_\tau$, and let (X_s, X_t) be an optimal coupling of (μ_s,μ_t), for the cost function $c^{s,t}$. Further let $(\gamma_\tau)_{s\le\tau\le t}$ be a random continuous path, such that law $(\gamma_\tau) = \mu_\tau$ for all $\tau\in[s,t]$. Then

$$\sum_i C^{t_i,t_{i+1}}(\mu_{t_i},\mu_{t_{i+1}}) \le C^{s,t}(\mu_s,\mu_t) = \mathbb{E}\,c^{s,t}(X_s,X_t) \le \mathbb{E}\,c^{s,t}(\gamma_s,\gamma_t) \le \mathbb{E}\,\mathcal{A}^{s,t}(\gamma),$$

where the but-to-last inequality follows from the fact that (γ_s,γ_t) is a coupling of (μ_s,μ_t), and the last inequality is a consequence of the definition of $c^{s,t}$. This shows that

$$\sum_i C^{t_i,t_{i+1}}(\mu_{t_i},\mu_{t_{i+1}}) \le \mathbb{E}\,\mathcal{A}^{s,t}(\gamma).$$

On the other hand, there is equality in the whole chain of inequalities if $t_0 = s$, $t_1 = t$, $X_s = \gamma_s$, $X_t = \gamma_t$, and γ_τ is a (random) action-minimizing curve. So the two expressions in (iii) do coincide.

Now let us address the equivalence between (ii) and (iii). First, it is clear that $\mathbb{A}^{s,t}$ is lower semi-continuous, since it is defined as a supremum of lower semi-continuous functionals. Then the inequality $\mathbb{A}^{t_1,t_3} \ge \mathbb{A}^{t_1,t_2} + \mathbb{A}^{t_2,t_3}$ holds true for all intermediate times

$t_1 < t_2 < t_3$ (this is a simple consequence of the definitions), and the converse inequality is a consequence of the general inequality

$$s < t_2 < t \Longrightarrow \quad C^{s,t_2}(\mu_s, \mu_{t_2}) \le C^{s,t_2}(\mu_s, \mu_{t_2}) + C^{t_2,t}(\mu_{t_2}, \mu_t),$$

which we proved in Step 1 above. So Property (i) in Definition 7.10 is satisfied. To check Property (ii) of that Definition, take any two probability measures μ_s, μ_t and introduce a displacement interpolation $(\mu_\tau)_{s\le\tau\le t}$ for the Lagrangian action restricted to $[s,t]$. Then Property (ii) of Theorem 7.19 implies that $\mathbb{A}^{s,t}(\mu) = C^{s,t}(\mu_s, \mu_t)$. Finally, Property (iii) in Definition 7.10 is also satisfied by construction. In the end, $(\mathbb{A})$ does define a Lagrangian action, with induced cost functionals $C^{s,t}$.

To conclude the proof of Theorem 7.19 it only remains to check the coercivity of the action; then the equivalence of (i), (ii) and (iii) will follow from Proposition 7.15(iv). Let $s < t$ be two given times in $[0,1]$, and let $\mathcal{K}_s, \mathcal{K}_t$ be compact sets of probability measures such that $C^{s,t}(\mu_s, \mu_t) < +\infty$ for all $\mu_s \in \mathcal{K}_s$, $\mu_t \in \mathcal{K}_t$. Action-minimizing curves for $\mathbb{A}^{s,t}$ can be written as law $(\gamma_\tau)_{s\le\tau\le t}$, where γ is a random action-minimizing curve $[s,t] \to \mathcal{X}$ such that law $(\gamma_s) \in \mathcal{K}_s$, law $(\gamma_t) \in \mathcal{K}_t$. One can use an argument similar to the one in Step 4 above to prove that the laws $\mathit{\Pi}$ of such minimizing curves form a tight, closed set; so we have a compact set of dynamical transference plans $\mathit{\Pi}^{s,t}$, that are probability measures on $C([s,t];\mathcal{X})$. The problem is to show that the paths $(e_\tau)_\# \mathit{\Pi}^{s,t}$ constitute a compact set in $C([s,t]; P(\mathcal{X}))$. Since the continuous image of a compact is compact, it suffices to check that the map

$$\mathit{\Pi}^{s,t} \longmapsto ((e_\tau)_\# \mathit{\Pi}^{s,t})_{s\le\tau\le t}$$

is continuous from $P(C([s,t];\mathcal{X}))$ to $C([s,t]; P(\mathcal{X}))$. To do so, it will be convenient to metrize $P(\mathcal{X})$ with the Wasserstein distance W_1, replacing if necessary d by a bounded, topologically equivalent distance. (Recall Corollary 6.10.) Then the uniform distance on $C([s,t];\mathcal{X})$ is also bounded and there is an associated Wasserstein distance $\mathcal{W}_1$ on $P(C([s,t];\mathcal{X}))$. Let $\mathit{\Pi}$ and $\widetilde{\mathit{\Pi}}$ be two dynamical optimal transference plans, and let $((\gamma_\tau), (\widetilde{\gamma}_\tau))$ be an optimal coupling of $\mathit{\Pi}$ and $\widetilde{\mathit{\Pi}}$; let also μ_τ, $\widetilde{\mu}_\tau$ be the associated displacement interpolations; then the required continuity follows from the chain of inequalities

$$\sup_{t\in[0,1]} W_1(\mu_t, \widetilde{\mu}_t) \le \sup_{t\in[0,1]} \mathbb{E}\, d(\gamma_t, \widetilde{\gamma}_t) \le \mathbb{E} \sup_{t\in[0,1]} d(\gamma_t, \widetilde{\gamma}_t) = \mathcal{W}_1(\mathit{\Pi}, \widetilde{\mathit{\Pi}}).$$

This proves that displacement interpolations with endpoints lying in given compact sets themselves form a compact set, and this concludes the proof of the coercivity of the action $(\mathbb{A}^{s,t})$. □

Remark 7.26. In the proof of the implication (ii) ⇒ (i), instead of defining $\mathit{\Pi}^{(k)}$ on the space of continuous curves, one could instead work with $\mathit{\Pi}^{(k)}$ defined on discrete times, construct by compactness a consistent system of marginals on $\mathcal{X}^{2^\ell+1}$, for all ℓ, and then invoke Kolmogorov's existence theorem to get a $\mathit{\Pi}$ which is defined on a set of curves. Things however are not that simple, since Kolmogorov's theorem constructs a random measurable curve which is not a priori continuous. Here one has the same conceptual difficulty as in the construction of Brownian motion as a probability measure on continuous paths.

Proof of Corollary 7.20. Introduce the family of actions

$$\mathcal{A}^{s,t}(\gamma) = \int_s^t |\dot{\gamma}_\tau|^p \, d\tau.$$

Then

$$c^{s,t}(x,y) = \frac{d(x,y)^p}{(t-s)^{p-1}},$$

and all our assumptions hold true for this action and cost. (The assumption of local compactness is used to prove that the action is coercive, see the Appendix.) The important point now is that

$$C^{s,t}(\mu,\nu) = \frac{W_p(\mu,\nu)^p}{(t-s)^{p-1}}.$$

So, according to the remarks in Example 7.8, Property (ii) in Theorem 7.19 means that (μ_t) is in turn a minimizer of the action associated with the Lagrangian $|\dot{\mu}|^p$, i.e. a geodesic in $P_p(\mathcal{X})$. Note that if μ_t is the law of a random optimal geodesic γ_t at time t, then

$$W_p(\mu_t,\mu_s)^p \le \mathbb{E}\, d(\gamma_s,\gamma_t)^p = \mathbb{E}\,(t-s)^p\, d(\gamma_0,\gamma_1)^p = (t-s)^p\, W_p(\mu_0,\mu_1)^p,$$

so the path μ_t is indeed continuous (and actually 1-Lipschitz) for the distance W_p. □

Proof of Corollary 7.21. By Theorem 7.19, any displacement interpolation has to take the form $(e_t)_\# \mathit{\Pi}$, where $\mathit{\Pi}$ is a probability measure on action-minimizing curves such that $\pi := (e_0, e_1)_\# \mathit{\Pi}$ is an optimal transference plan. By assumption, there is exactly one such π. Let Z be the set of couples (x_0, x_1) such that there is more than one minimizing curve joining x_0 to x_1; by assumption $\pi[Z] = 0$. For $(x_0, x_1) \notin Z$, there is a unique geodesic $\gamma = S(x_0, x_1)$ joining x_0 to x_1. So $\mathit{\Pi}$ has to coincide with $S_\#\pi$. □

Displacement interpolation between intermediate times

Let again μ_0 and μ_1 be any two probability measures, $(\mu_t)_{0\le t\le 1}$ a displacement interpolation associated with a dynamical optimal transference plan $\mathit{\Pi}$, and $(\gamma_t)_{0\le t\le 1}$ a random action-minimizing curve with law $(\gamma) = \mathit{\Pi}$. In particular (γ_0, γ_1) is an optimal coupling of (μ_0, μ_1). With the help of the previous arguments, it is almost obvious that $(\gamma_{t_0}, \gamma_{t_1})$ is also an optimal coupling of (μ_{t_0}, μ_{t_1}). What may look at first sight more surprising is that if $(t_0, t_1) \neq (0, 1)$, this is the *only* optimal coupling, at least if action-minimizing curves are not "branching". Furthermore, there is a restriction property, which generalizes Theorem 4.5 This is the content of the next theorem.

Theorem 7.27 (Interpolation from intermediate times). *Let $\mathcal{X}$ be a Polish space equipped with a coercive action $(\mathcal{A})$ on $C([0,1];\mathcal{X})$. Let $\mathit{\Pi} \in P(C([0,1];\mathcal{X}))$ be a dynamical optimal transport plan. For any t_0, t_1 in $[0,1]$ with $0 \le t_0 < t_1 \le 1$, define its time-restriction $\mathit{\Pi}^{t_0,t_1}$ as $(r_{t_0,t_1})_\# \mathit{\Pi}$, where $r_{t_0,t_1}(\gamma)$ is the restriction of γ to the interval $[t_0, t_1]$. Then,*

(i) $\mathit{\Pi}^{t_0,t_1}$ is a dynamical optimal coupling for the action $(\mathcal{A})^{t_0,t_1}$ on the time interval $[t_0, t_1]$;

(ii) If $\widetilde{\mathit{\Pi}}$ is a nonnegative measure on $C([t_0,t_1];\mathcal{X})$, such that $\widetilde{\mathit{\Pi}} \le \mathit{\Pi}^{t_0,t_1}$ and $\widetilde{\mathit{\Pi}}[C([t_0,t_1];\mathcal{X})] > 0$, then

$$\mathit{\Pi}' := \frac{\widetilde{\mathit{\Pi}}}{\widetilde{\mathit{\Pi}}\big[C([t_0,t_1];\mathcal{X})\big]}$$

is a dynamical optimal coupling between $\mu'_{t_0} := (e_{t_0})_\# \mathit{\Pi}'$ and $\mu'_{t_1} := (e_{t_1})_\# \mathit{\Pi}'$;

(iii) Assume further that action-minimizing curves are uniquely and measurably determined by their restriction to a nontrivial time-interval, and $(t_0, t_1) \neq (0, 1)$. Then,

Π' in (ii) is the unique dynamical optimal coupling between μ'_{t_0} and μ'_{t_1}. In particular, $(\pi')^{t_0,t_1} := (e_{t_0}, e_{t_1})_\# \Pi'$ is the unique optimal transference plan between μ'_{t_0} and μ'_{t_1}; and $\mu'_t := (e_t)_\# \Pi'$ $(t_0 \le t \le t_1)$ defines the unique displacement interpolation between μ'_{t_0} and μ'_{t_1}.

(iv) Under the same assumptions as (iii), for any $t \in (0,1)$, $(\Pi \otimes \Pi)(d\gamma\, d\widetilde{\gamma})$-almost surely,

$$[\gamma_t = \widetilde{\gamma}_t] \Longrightarrow [\gamma = \widetilde{\gamma}].$$

In other words, the curves seen by Π cannot cross at intermediate times. Moreover, one can define a measurable map $F_t : \mathcal{X} \to \Gamma(\mathcal{X})$ such that, Π-almost surely, $F_t(\gamma_t) = \gamma$.

Remark 7.28. In Chapter 8 we shall work in the setting of smooth Riemannian manifolds and derive a quantitative variant of the "no-crossing" property expressed in (iv).

Corollary 7.29. *Let $(\mathcal{X}, d)$ be a complete separable, locally compact length space. Let $p > 1$ and let $P_p(\mathcal{X})$ be the space of probability measures on $\mathcal{X}$ with finite moment of order p, metrized by the Wasserstein distance W_p. Assume that $\mathcal{X}$ is nonbranching, in the sense that a geodesic $\gamma : [0,1] \to \mathcal{X}$ is uniquely determined by its restriction to a nontrivial time-interval. Then also $P_p(\mathcal{X})$ is nonbranching. Conversely, if $P_p(\mathcal{X})$ is nonbranching, then $\mathcal{X}$ is nonbranching.*

Proof of Theorem 7.27. Let γ be a random geodesic with law Π. Let γ^{0,t_0}, γ^{t_0,t_1} and $\gamma^{t_1,1}$ stand for the restrictions of γ to the time intervals $[0,t_0]$, $[t_0,t_1]$ and $[t_1,1]$, respectively. Then

$$\begin{aligned} C^{0,t_0}(\mu_0,\mu_{t_0}) + C^{t_0,t_1}(\mu_{t_0},\mu_{t_1}) + C^{t_1,1}(\mu_{t_1},\mu_1) &\le \mathbb{E}\, c^{0,t_0}(\gamma^{0,t_0}) + \mathbb{E}\, c^{t_0,t_1}(\gamma^{t_0,t_1}) + \mathbb{E}\, c^{t_1,1}(\gamma^{t_1,1}) \\ &= \mathbb{E}\, c^{0,1}(\gamma) = C^{0,1}(\mu_0,\mu_1) \\ &\le C^{0,t_0}(\mu_0,\mu_{t_0}) + C^{t_0,t_1}(\mu_{t_0},\mu_{t_1}) + C^{t_1,1}(\mu_{t_1},\mu_1). \end{aligned}$$

So there has to be equality in all the inequalities, and it follows that

$$\mathbb{E}\, c^{t_0,t_1}(\gamma^{t_0},\gamma^{t_1}) = C^{t_0,t_1}(\mu_{t_0},\mu_{t_1}).$$

So γ^{t_0,t_1} is optimal, and Π^{t_0,t_1} is a dynamical optimal transference plan. Statement (i) is proven.

As a corollary of (i), $\pi^{t_0,t_1} = (e_{t_0}, e_{t_1})_\# \Pi^{t_0,t_1}$ is an optimal transference plan between μ_{t_0} and μ_{t_1}. Let $\widetilde{\pi} := (e_{t_0}, e_{t_1})_\# \widetilde{\Pi}$. The inequality $\widetilde{\Pi} \le \Pi^{t_0,t_1}$ is preserved by push-forward, so $\widetilde{\pi} \le \pi^{t_0,t_1}$. Also $\widetilde{\pi}[\mathcal{X} \times \mathcal{X}] = \widetilde{\Pi}[C([t_0,t_1];\mathcal{X})] > 0$. By Theorem 4.5, $\pi' := \widetilde{\pi}/\widetilde{\pi}[\mathcal{X} \times \mathcal{X}]$ is an optimal transference plan between its marginals. But π' coincides with $(e_{t_0}, e_{t_1})_\# \Pi'$, and since Π' is concentrated (just as Π) on action-minimizing curves, Theorem 7.19 guarantees that Π' is a dynamical optimal transference plan between its marginals. This proves (ii).

To prove (iii), assume, without loss of generality, that $t_0 > 0$; then an action-minimizing curve γ is uniquely and measurably determined by its restriction γ^{0,t_0} to $[0,t_0]$. In other words, there is a measurable function $F^{0,t_0} : \Gamma^{0,t_0} \to \Gamma$, defined on the set of all γ^{0,t_0}, such that any action-minimizing curve $\gamma : [0,1] \to \mathcal{X}$ can be written as $F^{0,t_0}(\gamma^{0,t_0})$. Similarly, there is a measurable function F^{t_0,t_1} such that $F^{t_0,t_1}(\gamma^{t_0,t_1}) = \gamma$.

By construction, $\widetilde{\Pi}$ is concentrated on the curves γ^{0,t_0}, that are the restrictions to $[t_0,t_1]$ of the action-minimizing curves γ. Let $\overline{\Pi} := (F^{t_0,t_1})_\# \widetilde{\Pi}$; this is a probability measure on $C([0,1];\mathcal{X})$. Of course $\overline{\Pi} \le (F^{t_0,t_1})_\# \Pi^{t_0,t_1} = \Pi$; so by (ii), $\overline{\Pi}/\overline{\Pi}[C([0,1];\mathcal{X})]$ is optimal. (In words, $\overline{\Pi}$ is obtained from Π by extending to the time-interval $[0,1]$ those curves

which appear in the sub-plan Π'.) Then it is easily seen that $\Pi' = (e_{t_0}, e_{t_1})_\# \overline{\Pi}$, and $\overline{\Pi}[C([0,1];\mathcal{X})] = \Pi'[C([t_0,t_1];\mathcal{X})]$. So it suffices to prove Theorem 7.27(iii) in the case when $\widetilde{\Pi} = \Pi^{t_0,t_1}$, and this will be assumed in the sequel.

Let now γ be a random geodesic with law Π, and $\Pi^{0,t_0} = \text{law}\,(\gamma^{0,t_0})$, $\Pi^{t_0,t_1} = \text{law}\,(\gamma^{t_0,t_1})$, $\Pi^{t_1,1} = \text{law}\,(\gamma^{t_1,1})$. By (i), Π^{t_0,t_1} is a dynamical optimal transference plan between μ_{t_0} and μ_{t_1}; let $\widetilde{\Pi}^{t_0,t_1}$ be another such plan. The goal is to show that $\widetilde{\Pi}^{t_0,t_1} = \Pi^{t_0,t_1}$.

Disintegrate Π^{0,t_0} and $\widetilde{\Pi}^{t_0,t_1}$ along their common marginal μ_{t_0} and glue them together. This gives a probability measure on $C([0,t_0];\mathcal{X}) \times \mathcal{X} \times C((t_0,t_1];\mathcal{X})$, supported on triples $(\gamma, g, \widetilde{\gamma})$ such that $\gamma(t) \to g$ as $t \to t_0^-$, $\widetilde{\gamma}(t) \to g$ as $t \to t_0^+$. Such triples can be identified with continuous functions on $[0,t_1]$, so what we have is in fact a probability measure on $C([0,t_1];\mathcal{X})$. Repeat the operation by gluing this with $\Pi^{t_1,1}$, so as to get a probability measure $\overline{\Pi}$ on $C([0,1];\mathcal{X})$.

Let then $\overline{\gamma}$ be a random variable with law $\overline{\Pi}$: By construction, law $(\overline{\gamma}^{0,t_0}) = \text{law}\,(\gamma^{0,t_0})$, and law $(\overline{\gamma}^{t_1,1}) = \text{law}\,(\gamma^{t_1,1})$, so

$$\begin{aligned}
\mathbb{E}\, c^{0,1}(\overline{\gamma}) &\le \mathbb{E}\, c^{0,t_0}(\overline{\gamma}^{0,t_0}) + \mathbb{E}\, c^{t_0,t_1}(\overline{\gamma}^{t_0,t_1}) + \mathbb{E}\, c^{t_1,1}(\overline{\gamma}^{t_1,1}) \\
&= \mathbb{E}\, c^{0,t_0}(\gamma^{0,t_0}) + \mathbb{E}\, c^{t_0,t_1}(\overline{\gamma}^{t_0,t_1}) + \mathbb{E}\, c^{t_1,1}(\gamma^{t_1,1}) \\
&= C^{0,t_0}(\mu_0,\mu_{t_0}) + C^{t_0,t_1}(\mu_{t_0},\mu_{t_1}) + C^{t_1,1}(\mu_{t_1},\mu_1) = C^{0,1}(\mu_0,\mu_1).
\end{aligned}$$

So $\overline{\Pi}$ is a dynamical optimal transference plan between μ_0 and μ_1. It follows from Theorem 7.19 that there is a random action-minimizing curve $\widehat{\gamma}$ with law $(\widehat{\gamma}) = \overline{\Pi}$. In particular,

$$\text{law}\,(\widehat{\gamma}^{0,t_0}) = \text{law}\,(\gamma^{0,t_0}); \qquad \text{law}\,(\widehat{\gamma}^{t_0,t_1}) = \text{law}\,(\widetilde{\gamma}^{t_0,t_1}).$$

But by assumption there is a measurable function F ($F = r_{t_0,t_1} \circ F^{0,t_0}$ such that $g^{t_0,t_1} = F(g^{0,t_0})$, for any action-minimizing curve g. Then

$$\text{law}\,(\widetilde{\gamma}^{t_0,t_1}) = \text{law}\,(\widehat{\gamma}^{t_0,t_1}) = \text{law}\,(F(\widehat{\gamma}^{0,t_0})) = \text{law}\,(F(\gamma^{0,t_0})) = \text{law}\,(\gamma^{t_0,t_1}).$$

This proves the uniqueness of the dynamical optimal transference plan joining μ'_{t_0} to μ'_{t_1}. The remaining part of (iii) is obvious since any optimal plan or displacement interpolation has to come from a dynamical optimal transference plan, according to Theorem 7.19.

Now we turn to the proof of (iv). Since $\pi = (e_0, e_1)_\# \Pi$ is cyclically monotone (Theorem 5.9), we have, $\Pi \otimes \Pi(d\gamma\, d\widetilde{\gamma})$-almost surely,

$$c^{0,1}(\gamma_0,\gamma_1) + c^{0,1}(\widetilde{\gamma}_0,\widetilde{\gamma}_1) \le c^{0,1}(\gamma_0,\widetilde{\gamma}_1) + c^{0,1}(\widetilde{\gamma}_0,\gamma_1). \tag{7.21}$$

If γ and $\widetilde{\gamma}$ are two such paths, assume that $\gamma_t = \widetilde{\gamma}_t = X$ for some $t \in (0,1)$. Then

$$c^{0,1}(\gamma_0,\widetilde{\gamma}_1) \le c^{0,t}(\gamma_0,X) + c^{t,1}(X,\widetilde{\gamma}_1), \tag{7.22}$$

and similarly

$$c^{0,1}(\widetilde{\gamma}_0,\gamma_1) \le c^{0,t}(\widetilde{\gamma}_0,X) + c^{t,1}(X,\gamma_1). \tag{7.23}$$

By adding up (7.22) and (7.23), we get

$$\begin{aligned}
c^{0,1}(\gamma_0,\widetilde{\gamma}_1) + c^{0,1}(\widetilde{\gamma}_0,\gamma_1) &\le \big[c^{0,t}(\gamma_0,X) + c^{t,1}(X,\gamma_1)\big] \,+\, \big[c^{0,t}(\widetilde{\gamma}_0,X) + c^{t,1}(X,\widetilde{\gamma}_1)\big] \\
&= c^{0,1}(\gamma_0,\gamma_1) + c^{0,1}(\widetilde{\gamma}_0,\widetilde{\gamma}_1).
\end{aligned}$$

Since the reverse inequality holds true by (7.21), equality has to hold in all intermediate inequalities, for instance in (7.22). Then it is easy to see that the path $\overline{\gamma}$ defined by $\overline{\gamma}(s) = \gamma(s)$ for $0 \le s \le t$, and $\overline{\gamma}(s) = \widetilde{\gamma}(s)$ for $s \ge t$, is a minimizing curve. Since it

coincides with γ on a nontrivial time-interval, it has to coincide with γ everywhere, and similarly it has to coincide with $\widetilde{\gamma}$ everywhere. So $\gamma = \widetilde{\gamma}$.

The previous conclusion holds true not only $\Pi \otimes \Pi$ almost surely, but actually for any two minimizing curves γ, $\widetilde{\gamma}$ which lie in the support of Π. Indeed, inequality (7.21) defines a closed set $\mathcal{C}$ in $\Gamma \times \Gamma$, where Γ stands for the set of minimizing curves; so $\mathrm{Spt}\,\Pi \times \mathrm{Spt}\,\Pi = \mathrm{Spt}(\Pi \otimes \Pi) \subset \mathcal{C}$.

Now let $(K_\ell)_{\ell \in \mathbb{N}}$ be a nondecreasing sequence of compact sets contained in $\mathrm{Spt}\,\Pi$, such that $\Pi[\cup K_\ell] = 1$. For each ℓ, we define F_ℓ on $e_t(K_\ell)$ by $F_\ell(\gamma_t) = \gamma$. This map is continuous: Indeed, if $x_k \in e_t(K_\ell)$ converges to x, then we may always assume that $x_k = (\gamma_k)_t$ for some $\gamma_k \in K_\ell$, and then up to extraction γ_k converges to $\gamma \in K_\ell$, and in particular $(\gamma_k)_t$ converges to γ_t, but then $F_\ell(\gamma_t) = \gamma$. Then we can define F on $\cup K_\ell$ as a map which coincides with F_ℓ on each K_ℓ. (Obviously, this is the same line of reasoning as in Theorem 5.25.) □

Proof of Corollary 7.29. Assume that $\mathcal{X}$ is nonbranching. Then there exists some function F, defined on the set of all functions γ^{0,t_0}, where γ^{0,t_0} is the restriction to $[0,t_0]$ of the geodesic $\gamma : [0,1] \to \mathcal{X}$, such that $\gamma = F(\gamma^{0,t_0})$.

Then F is automatically continuous. Indeed, let $(\gamma_n)_{n\in\mathbb{N}}$ be such that γ_n^{0,t_0} converges uniformly on $C([0,t_0];\mathcal{X})$ to some curve $g : [0,t_0] \to \mathcal{X}$. Since the functions γ_n^{0,t_0} are uniformly bounded, the speeds of all the geodesics γ_n are uniformly bounded too, and the curves $\gamma_n([0,1])$ are all included in a compact subset of $\mathcal{X}$. It follows from Ascoli's theorem that the sequence (γ_n) converges uniformly, up to extraction of a subsequence. But then its limit γ has to be a geodesic, and its restriction to $[0,t_0]$ should coincide with g. There is at most one such geodesic, so γ is uniquely determined, and the whole sequence γ_n converges. This implies the continuity of F.

Then Theorem 7.27(iii) applies: If $(\mu_t)_{0\le t\le 1}$ is a geodesic in $P_p(\mathcal{X})$, then there is only one geodesic between μ_{t_0} and μ_1. So $(\mu_t)_{0\le t\le 1}$ is uniquely determined by its restriction to $[0,t_0]$. The same reasoning could be done for any nontrivial time-interval instead of $[0,t_0]$; so $P_p(\mathcal{X})$ is indeed nonbranching.

The converse implication is obvious, since any geodesic γ in $\mathcal{X}$ induces a geodesic in $P_p(\mathcal{X})$, namely $(\delta_{\gamma(t)})_{0\le t\le 1}$. □

Interpolation of prices

When the path μ_t varies in time, one can ask what becomes of the pair of "prices" (ψ,ϕ) in the Kantorovich duality? The short answer is that these functions will also evolve continuously in time, according to **Hamilton–Jacobi equations**.

Definition 7.30 (Hamilton–Jacobi–Hopf–Lax–Oleinik semigroup). *Let $\mathcal{X}$ be a metric space and $(\mathcal{A})^{0,1}$ a coercive Lagrangian action on $\mathcal{X}$, with continuous cost functions $(c^{s,t})_{0\le s<t\le 1}$. For any two measurable functions $\psi : \mathcal{X} \to \mathbb{R}\cup\{+\infty\}$, $\phi : \mathcal{X} \to \mathbb{R}\cup\{-\infty\}$, and any two times $0 \le s < t \le 1$, define*

$$\begin{cases} H_+^{s,t}\psi\,(y) = \inf\limits_{x\in\mathcal{X}} \Big(\psi(x) + c^{s,t}(x,y)\Big); \\ \\ H_-^{t,s}\phi\,(x) = \sup\limits_{y\in\mathcal{X}} \Big(\phi(y) - c^{s,t}(x,y)\Big). \end{cases}$$

The family of operators $(H_+^{s,t})_{t>s}$ (resp. $(H_-^{s,t})_{s<t}$) is called the forward (resp. backward) Hamilton–Jacobi (or Hopf–Lax, or Lax–Oleinik) semigroup.

Roughly speaking, $H_+^{s,t}$ gives the values of ψ at time t, from its values at time s; while $H_-^{s,t}$ does the reverse. So the semigroups H_- and H_+ are in some sense inverse of each other. Yet it is not true in general that $H_-^{s,t} H_+^{s,t} = \mathrm{Id}$. Proposition 7.31 below summarizes some of the main properties of these semigroups; the denomination of "semigroup" itself is justified by Property (ii).

Proposition 7.31. *With the notation of Definition 7.30,*

(i) $H_+^{s,t}$ *and* $H_-^{s,t}$ *are order-preserving, that is* $\psi \le \overline{\psi} \Longrightarrow H_\pm^{s,t}\psi \le H_\pm^{s,t}\overline{\psi}$.

(ii) Whenever $t_1 < t_2 < t_3$ *are three intermediate times in* $[0,1]$,

$$\begin{cases} H_+^{t_2,t_3} H_+^{t_1,t_2} = H_+^{t_1,t_3} \\ \\ H_-^{t_2,t_1} H_-^{t_3,t_2} = H_-^{t_3,t_1}. \end{cases}$$

(iii) Whenever $s < t$ *are two times in* $[0,1]$,

$$H_-^{t,s} H_+^{s,t} \le \mathrm{Id}\,; \qquad H_+^{s,t} H_-^{t,s} \ge \mathrm{Id}\,.$$

Proof. Properties (i) and (ii) are immediate consequences of the definitions and Proposition 7.15(iii). To check Property (iii), e.g. the first half of it, write

$$H_-^{t,s}(H_+^{s,t}\psi)(x) = \sup_y \inf_{x'} \Big(\psi(x') + c^{s,t}(x',y) - c^{s,t}(x,y)\Big).$$

The choice $x' = x$ shows that the infimum above is bounded above by $\psi(x)$, independently of y; so $H_-^{t,s}(H_+^{s,t}\psi)(x) \le \psi(x)$, as desired. □

The Hamilton–Jacobi semigroup is well-known and useful in geometry and dynamical systems theory. On a smooth Riemannian manifold, when the action is given by a Lagrangian $L(x,v,t)$, strictly convex and superlinear in the velocity variable, then $S_+(t,\cdot) := H_+^{0,t}\psi_0$ solves the differential equation

$$\frac{\partial S_+}{\partial t}(x,t) + H\big(x, \nabla S_+(x,t), t\big) = 0, \tag{7.24}$$

where $H = L^*$ is obtained from L by Legendre transform in the v variable, and is called the **Hamiltonian** of the system. This equation provides a bridge between a Lagrangian description of action-minimizing curves, and an Eulerian description: From $S_+(x,t)$ one can reconstruct a velocity field $sv(x,t) = \nabla_p H\big(x, \nabla S_+(x,t), t\big)$, in such a way that integral curves of the equation $\dot{x} = v(x,t)$ are minimizing curves. Well, rigorously speaking, that would be the case if S_+ were differentiable! But things are not so simple because S_+ is not in general differentiable everywhere, so the equation has to be interpreted in a suitable sense (called viscosity sense). It is important to note that if one uses the backward semigroup and defines $S_-(x,t) := H_-^{t,1}\psi_t$, then S_- formally satisfies the same equation as S_+, but the equation has to be interpreted with a different convention (backward viscosity). This will be illustrated by the next example.

Example 7.32. On a smooth Riemannian manifold M, consider the simple Lagrangian cost $L(x,v,t) = |v|^2/2$; then the associated Hamiltonian is just $H(x,p,t) = |p|^2/2$. If S is a C^1 solution of $\partial S/\partial t + |\nabla S|^2/2 = 0$, then the gradient of S can be interpreted as the velocity field of a family of minimizing geodesics. But if S_0 is a given Lipschitz function

and $S_+(t,x)$ is defined by the forward Hamilton–Jacobi semigroup starting from initial datum S_0, one only has (for all t, x)

$$\frac{\partial S_+}{\partial t} + \frac{|\nabla^- S_+|^2}{2} = 0,$$

where

$$|\nabla^- f|(x) := \limsup_{y \to x} \frac{[f(y) - f(x)]_-}{d(x,y)}, \qquad z_- = \max(-z, 0).$$

Conversely, if one uses the backward Hamilton–Jacobi semigroup to define a function $S_-(x,t)$, then

$$\frac{\partial S_-}{\partial t} + \frac{|\nabla^+ S_-|^2}{2} = 0, \qquad |\nabla^+ f|(x) := \limsup_{y \to x} \frac{[f(y) - f(x)]_+}{d(x,y)}, \quad z_+ = \max(z, 0).$$

When the Lagrangian is more complicated, things may become much more intricate. The standard convention is to use the forward Hamilton–Jacobi semigroup by default.

Now we shall see that the Hamilton–Jacobi semigroup provides a simple answer to the problem of interpolation in dual variables. In the next statement, $\mathcal{X}$ is again a Polish space, $(\mathcal{A})^{0,1}$ a coercive Lagrangian action on $\mathcal{X}$, with associated cost functions $c^{s,t}$; and $C^{s,t}$ stands for the optimal total cost in the transport problem with cost $c^{s,t}$.

Theorem 7.33 (Interpolation of prices). *With the same assumptions and notation as in Definition 7.30, let μ_0, μ_1 be two probability measures on $\mathcal{X}$, such that $C^{0,1}(\mu_0, \mu_1) < +\infty$, and let (ψ_0, ϕ_1) be a pair of $c^{0,1}$-conjugate functions such that any optimal plan $\pi_{0,1}$ between μ_0 and μ_1 has its support included in $\partial_{c^{0,1}} \psi_0$. (Recall Theorem 5.9; under adequate integrability conditions, the pair (ψ_0, ϕ_1) is just a solution of the dual Kantorovich problem.) Let further $(\mu_t)_{0 \le t \le 1}$ be a displacement interpolation between μ_0 and μ_1. Whenever $s < t$ are two intermediate times in $[0,1]$, define*

$$\psi_s := H_+^{0,s} \psi_0, \qquad \phi_t := H_-^{t,1} \phi_1.$$

Then (ψ_s, ϕ_t) is optimal in the dual Kantorovich problem associated to (μ_s, μ_t) and $c^{s,t}$. In particular,

$$C^{s,t}(\mu_s, \mu_t) = \int \phi_t \, d\nu_t - \int \psi_s \, d\mu_s,$$

and

$$\phi_t(y) - \psi_s(x) \le c^{s,t}(x,y),$$

with equality $(\mu_s \otimes \mu_t)(dx\, dy)$-almost surely.

Proof. From the definitions,

$$\phi_t(y) - \psi_s(x) - c^{s,t}(x,y) = \sup_{y', x'} \Big[\phi_1(y') - c^{t,1}(y', y) - \psi_0(x') - c^{0,s}(x', x) - c^{s,t}(x,y) \Big].$$

Since $c^{0,s}(x', x) + c^{s,t}(x,y) + c^{t,1}(y, y') \ge c^{0,1}(x', y')$, it follows that

$$\phi_t(y) - \psi_s(x) - c^{s,t}(x,y) \le \sup_{y', x'} \Big[\phi_1(y') - \psi_0(x') - c^{0,1}(x', y') \Big] \le 0.$$

So $\phi_t(y) - \psi_s(x) \le c^{s,t}(x,y)$. This inequality does not depend on the fact that (ψ_0, ϕ_1) is a tight pair of prices, in the sense of (5.5), but only on the inequality $\phi_1 - \psi_0 \le c^{0,1}$.

Next, introduce a random action-minimizing curve γ such that $\mu_t = \text{law}\,(\gamma_t)$. Since (ψ_0, ϕ_1) is an optimal pair, we know from Theorem 5.9 that, almost surely,

$$\phi_1(\gamma_1) - \psi_0(\gamma_0) = c^{0,1}(\gamma_0, \gamma_1).$$

From the identity $c^{0,1}(\gamma_0, \gamma_1) = c^{0,s}(\gamma_0, \gamma_s) + c^{s,t}(\gamma_s, \gamma_t) + c^{t,1}(\gamma_t, \gamma_1)$ and the definition of ψ_s and ϕ_t, it follows that

$$c^{s,t}(\gamma_s, \gamma_t) = \big[\phi_1(\gamma_1) - c^{t,1}(\gamma_t, \gamma_1)\big] - \big[\psi_0(\gamma_0) + c^{0,s}(\gamma_0, \gamma_s)\big] \leq \phi_t(\gamma_t) - \psi_s(\gamma_s).$$

This shows that actually $c^{s,t}(\gamma_s, \gamma_t) = \phi_t(\gamma_t) - \psi_s(\gamma_s)$ almost surely, so (ψ_s, ϕ_t) has to be optimal in the dual Kantorovich problem from $\mu_s = \text{law}\,(\gamma_s)$ to $\mu_t = \text{law}\,(\gamma_t)$. □

Remark 7.34. In the limit case $s \to t$, the above results become

$$\begin{cases} \phi_t \leq \psi_t \\ \phi_t = \psi_t \quad \mu_t\text{-almost surely} \end{cases}$$

... but it is not true in general that $\phi_t = \psi_t$ everywhere in $\mathcal{X}$.

Exercise 7.35. After reading the rest of Part I, the reader can come back to this exercise and check his or her understanding by proving that, for a quadratic Lagrangian,

(i) The displacement interpolation between two balls in Euclidean space is always a ball, with linearly increasing radius (here I am identifying a set with the uniform probability measure on this set);

(ii) More generally, the displacement interpolation between two ellipsoids is always an ellipsoid;

(iii) But the displacement interpolation between two sets is in general not a set;

(iv) The displacement interpolation between two spherical caps on the sphere is in general not a spherical cap;

(v) The displacement interpolation between two antipodal spherical caps on the sphere is unique, while the displacement interpolation between two antipodal points can be realized in an infinite number of ways.

Appendix: Paths in metric structures

This Appendix is a kind of crash basic course in Riemannian geometry, and nonsmooth generalizations thereof. Much more details can be obtained from the references cited in the bibliographical notes.

A Riemannian manifold is a (finite-dimensional) manifold M equipped with a **metric** g: this means that g defines a scalar product on each tangent space $T_x M$, varying smoothly with x. So if v and w at tangent vectors at x, the notation $v \cdot w$ really means $g_x(v, w)$, where g_x is the metric at x. The degree of smoothness of g depends on the context, but it is customary to consider C^3 manifolds with a C^2 metric. For the purpose of these notes, the reader might assume C^∞ smoothness.

Let $\gamma : [0, 1] \to M$ be a smooth path, denoted $(\gamma_t)_{0 \leq t \leq 1}$. (For me the words "path" and "curve" are synonymous.) For each $t \in (0, 1)$, the time-derivative at time t is — by the very definition of tangent space — the tangent vector $v = \dot{\gamma}_t$ in $T_{\gamma_t} M$. The scalar product

g gives a way to measure the norm of that vector: $|v|_{T_xM} = \sqrt{v \cdot v}$. Then one can define the **length** of γ by the formula

$$\mathcal{L}(\gamma) = \int_0^1 |\dot\gamma_t|\, dt, \tag{7.25}$$

and the **distance**, or geodesic distance, between two points x and y by the formula

$$d(x,y) = \inf \Big\{ \mathcal{L}(\gamma); \quad \gamma_0 = x, \ \gamma_1 = y \Big\}. \tag{7.26}$$

After that it is easy to extend the length formula to absolutely continuous curves. Note that any one of the three objects (metric, length, distance) determines the other two; indeed, the metric can be recovered from the distance via the formula

$$|\dot\gamma_0| = \lim_{t \downarrow 0} \frac{d(\gamma_0, \gamma_t)}{t}, \tag{7.27}$$

and the usual polarization identity

$$g(v,w) = \frac{1}{4} \Big[g(v+w, v+w) - g(v-w, v-w) \Big].$$

Let TM stand for the collection of all T_xM, $x \in M$, equipped with a manifold structure which is locally product. A point in TM is a couple (x,v) with $v \in T_xM$. The map $\pi : (x,v) \longmapsto x$ is the projection of TM onto M; it is obviously smooth and surjective. A function $M \to TM$ is called a vector field: It is given by a tangent vector at each point. So a vector field really is a map $f : x \to (x,v)$, but by abuse of notation one often writes $f(x) = v$. If $\gamma : [0,1] \to M$ is an injective path, one defines a *vector field along* γ as a path $\xi : [0,1] \to TM$ such that $\pi \circ \xi = \gamma$.

If $p = p(x)$ is a linear form varying smoothly on T_xM, then it can be identified, thanks to the metric g, to a vector field ξ, via the formula

$$p(x) \cdot v = \xi(x) \cdot v,$$

where $v \in T_xM$, and the dot in the left-hand side just means "$p(x)$ applied to v", while the dot in the right-hand side stands for the scalar product defined by g. As a particular case, if p is the differential of a function f, the corresponding vector field ξ is the *gradient* of f, denoted by ∇f or $\nabla_x f$.

If $f = f(x,v)$ is a function on TM, then one can differentiate it with respect to x or with respect to v. Since $T_(x,v)T_xM \simeq T_xM$, both $d_x f$ and $d_v f$ can be seen as linear forms on T_xM; this allows to define $\nabla_x f$ and $\nabla_v f$, the "gradient with respect to the position variable" and the "gradient with respect to the velocity variable".

Differentiating functions does not pose any particular conceptual problem, but differentiating vector fields is quite a different story. If ξ is a vector field on M, then $\xi(x)$ and $\xi(y)$ live in different vector spaces, so it does not a priori make sense to compare them, unless one *identifies* in some sense T_xM and T_yM. (Of course, one could say that ξ is a map $M \to TM$ and define its differential as a map $TM \to T(TM)$ but this is of little use, because $T(TM)$ is "too large"; it is much better if we can come up with a reasonable notion of derivation which produces a map $TM \to TM$.)

There is in general no canonical way to identify T_xM and T_yM if $x \neq y$; but there is a canonical way to identify $T_{\gamma_0}M$ and $T_{\gamma_t}M$ as t varies continuously. This operation is called *parallel transport*, or Levi-Civita transport. A vector field which is transported in

a parallel way along the curve γ will "look constant" to an observer who lives in M and travels along γ. If M is a surface embedded in $\mathbb{R}^3$, parallel transport can be described as follows: Start from the tangent plane at $\gamma(0)$, and then press your plane onto M along γ, in such a way that there is no friction (no slip, no rotation) between the plane and the surface.

With this notion it becomes possible to compute the derivative of a vector field *along a path*: If γ is a path and ξ is a vector field along γ, then the derivative of ξ is the another vector field along γ, say $t \to \dot{\xi}(t)$, defined by

$$\dot{\xi}(t_0) = \frac{d}{dt}\bigg|_{t=t_0} \theta_{t\to t_0}(\xi(\gamma_t)),$$

where $\theta_{t\to t_0}$ is the parallel transport from $T_{\gamma_t}M$ to $T_{\gamma_{t_0}}M$ along γ. This makes sense because $\theta_{t\to t_0}\xi(\gamma_t)$ is an element of the fixed vector space $T_{\gamma_{t_0}}M$. The path $t \to \dot{\xi}(t)$ is a vector field along γ, called the **covariant derivative** of ξ along γ, and denoted by $\nabla_{\dot\gamma}\xi$, or, if there is no possible confusion about the choice of γ, $D\xi/Dt$ (or simply $d\xi/dt$). If $M = \mathbb{R}^n$, then $\nabla_{\dot\gamma}\xi$ coincides with $(\dot\gamma \cdot \nabla)\xi$.

It turns out that the value of $\dot{\xi}(t_0)$ only depends on γ_{t_0}, on the values of ξ in a neighborhood of γ_{t_0}, and on the velocity $\dot\gamma_{t_0}$ (not on the whole path γ_t). Thus if ξ is a vector field defined in the neighborhood of a point x, and v is a tangent vector at x, it makes sense to define $\nabla_v\xi$ by the formula

$$\nabla_v\xi(x) = \frac{D\xi}{Dt}(\gamma_0), \qquad \gamma_0 = x, \quad \dot\gamma_0 = v.$$

The quantity $\nabla_v\xi(x)$ is "the covariant derivative of the vector field ξ at x in the direction v." Of course, if ξ and v are two vector fields, one can define a vector field $\nabla_v\xi$ by the formula $(\nabla_v\xi)(x) = (\nabla_{v(x)}\xi)(x)$. The linear operator $v \to \nabla_v\xi(x)$ is the **covariant gradient** of ξ at x; it is a linear operation $T_xM \to T_xM$.

It is worth noticing explicitly that the notion of covariant derivation coincides with the *convective* derivation used in fluid mechanics (for instance in Euler's equation for an incompressible fluid). I shall sometimes adopt the notation classically used in fluid mechanics: $(\nabla\xi)v = v \cdot \nabla\xi$. (On the contrary, the notation $(\nabla\xi) \cdot v$ should rather be reserved for $(\nabla\xi)^* v$, where $(\nabla\xi)^*$ is the adjoint of $\nabla\xi$; then $\langle v \cdot \nabla\xi, w\rangle = \langle v, \nabla\xi \cdot w\rangle$ and we are back to the classical conventions of $\mathbb{R}^n$.)

The procedure of parallel transport allows one to define the covariant derivation; conversely, the equations of parallel transport along γ can be written as $D\xi/Dt = 0$, where D/Dt is the covariant derivative along γ. So it is equivalent to define the notion of covariant derivation, or to define the rules of parallel transport.

There are (at least) three points of view about the covariant derivation. The first one is the *extrinsic* point of view: Let us think of M as an embedded surface in $\mathbb{R}^N$; that is, M is a subset of $\mathbb{R}^N$, it is equipped with the topology induced by $\mathbb{R}^N$, and the quadratic form g_x is just the usual Euclidean scalar product on $\mathbb{R}^N$, restricted to T_xM. Then the covariant derivative is just defined by

$$\dot{\xi}(t) = \Pi_{T_{\gamma_t}M}\left(\frac{d(\xi(\gamma_t))}{dt}\right),$$

where Π_{T_xM} stands for the orthogonal projection (in $\mathbb{R}^N$) onto T_xM. In short, the covariant derivative is the projection of the usual derivative onto the tangent space.

While this definition is very simple, it does not reveal the fact that the covariant derivation and parallel transport are *intrinsic* notions, which are invariant under isometry

and do not depend on the embedding of M into $\mathbb{R}^N$, but just on g. An intrinsic way to define covariant derivation is as follows: It is uniquely characterized by the two natural rules

$$\frac{d}{dt}\langle \xi,\, \zeta\rangle = \langle \dot\xi,\, \zeta\rangle + \langle \xi,\, \dot\zeta\rangle; \qquad \frac{D}{Dt}(f\xi) = \dot f\, \xi + f\, \dot\xi, \tag{7.28}$$

where the dependence of all the expressions on t is implicit; and by the not so natural rule

$$\nabla_\zeta \xi - \nabla_\xi \zeta = [\xi, \zeta].$$

Here $[\xi, \zeta]$ is the Lie bracket of ξ and ζ, which is defined as the unique vector field such that for any function F,

$$(dF)\cdot[\xi,\zeta] = d(dF\cdot\xi)\cdot\zeta - d(dF\cdot\zeta)\cdot\xi.$$

Further note that in the second formula of (7.28) the symbol $\dot f$ stands for the usual derivative of $t \to f(\gamma_t)$; while the symbols $\dot\xi$ and $\dot\zeta$ stand for the covariant derivatives of the vector fields ξ and ζ along γ.

A third point of view on covariant derivation is based on **coordinates**. Let $x \in M$, then there is a neighborhood O of x which is diffeomorphic to some open subset $U \subset \mathbb{R}^n$. Let Φ be a diffeomorphism $U \to O$, and let $(e_1, \ldots, e_n)$ be the usual basis of $\mathbb{R}^n$. Then a point m in O is said to have *coordinates* $(y^1, \ldots, y^n)$ if $m = \Phi(y^1, \ldots, y^n)$; and a vector $v \in T_m M$ is said to have *components* $v^1, \ldots, v^k$ if $d_{(y^1,\ldots,y^n)}\Phi \cdot (v_1, \ldots, v_k) = v$. Then the *coefficients* of the metric g are the functions g_{ij} defined by $g(v,v) = \sum g_{ij} v^i v^j$.

The coordinate point of view reduces everything to "explicit" computations and formulas in $\mathbb{R}^n$; for instance the derivation of a function f along the i^{th} direction is defined as $\partial_i f := (\partial/\partial y^i)(f \circ \Phi)$. This point of view is conceptually simple, but rapidly leads to cumbersome expressions. A central role in these formulas is played by the *Christoffel symbols*, which are defined by

$$\Gamma_{ij}^k := \frac{1}{2}\sum_{k=1}^n \Big(\partial_i g_{jk} + \partial_j g_{ki} - \partial_k g_{ij}\Big)\, g^{km},$$

where (g^{ij}) is by convention the inverse of (g_{ij}). Then the covariant derivation along γ is given by the formula

$$\left(\frac{D\xi}{Dt}\right)^k = \frac{d\xi^k}{dt} + \sum_{ij} \Gamma_{ij}^k\, \dot\gamma^i\, \xi^j.$$

Be it in the extrinsic or the intrinsic or the coordinate point of view, the notion of covariant derivative is one of the cornerstones on which differential Riemannian geometry has been constructed.

Another important concept is that of **Riemannian volume**, which I shall denote by vol. It can be defined intrinsically as the n-dimensional Hausdorff measure associated with the geodesic distance (where n is the dimension of the manifold). In coordinates, $\mathrm{vol}\,(dx) = \sqrt{\det(g)}\, dx$. The Riemannian volume plays the same role as the Lebesgue measure in $\mathbb{R}^n$.

After these reminders about Riemannian calculus, we can go back to the study of action minimization. Let $L(x,v,t)$ be a smooth Lagrangian on $TM \times [0,1]$. To find an equation satisfied by the curves which minimize the action, we can compute the differential of the action. So let γ be a curve, and h a small variation of that curve. (This amounts to considering a family $\gamma_{s,t}$ in such a way that $\gamma_{0,t} = \gamma_t$ and $(d/ds)\gamma_{s,t} = h(t)$.) Then the infinitesimal variation of the action $\mathcal{A}$ at γ, along the variation h, is

$$d\mathcal{A}(\gamma)\cdot h = \int_0^1 \Big(\nabla_x L(\gamma_t,\dot\gamma_t,t)\cdot h(t) + \nabla_v L(\gamma_t,\dot\gamma_t,t)\cdot \dot h(t)\Big)\,dt.$$

Thanks to (7.28) we can make an integration by parts with respect to the time variable, and get

$$d\mathcal{A}(\gamma)\cdot h = \int_0^1 \left(\nabla_x L - \frac{d}{dt}(\nabla_v L)\right)(\gamma_t,\dot\gamma_t,t)\ \cdot h(t)\,dt \\ + (\nabla_x L)(\gamma_1,\dot\gamma_1,1)\ \cdot h(1) - (\nabla_x L)(\gamma_0,\dot\gamma_0,0)\ \cdot h(0). \quad (7.29)$$

This is the **first variation formula**.

When the endpoints x, y of γ are fixed, then the tangent curve h vanishes at $t = 0$ and $t = 1$. Since h is otherwise arbitrary, it is easy to deduce the equation for minimizers:

$$\frac{d}{dt}\nabla_v L = \nabla_x L. \quad (7.30)$$

More explicitly, if a differentiable curve $(\gamma_t)_{0\le t\le 1}$ is minimizing, then

$$\frac{d}{dt}\Big(\nabla_v L(\gamma_t,\dot\gamma_t,t)\Big) = \nabla_x L(\gamma_t,\dot\gamma_t,t), \qquad 0 < t < 1.$$

This is the **Euler-Lagrange equation** associated with the Lagrangian L; to memorize it, it is convenient to write it as

$$\frac{d}{dt}\frac{\partial L}{\partial \dot x} = \frac{\partial L}{\partial x}, \quad (7.31)$$

so that the two time-derivatives in the left-hand side formally "cancel out". Note carefully that the left-hand side of the Euler-Lagrange equation involves the time-derivative of a curve which is valued in TM; so (d/dt) in (7.31) is in fact a covariant derivative along the minimizing curve γ, the same operation as we denoted before by $\nabla_{\dot\gamma}$, or D/Dt.

The most basic example is when $L(x,v,t) = |v|^2/2$. Then $\nabla_v L = v$ and the equation reduces to $dv/dt = 0$, or $\nabla_{\dot\gamma}\dot\gamma = 0$, which is the usual equation of vanishing acceleration. Curves with zero acceleration are called **geodesics**; their equation, is coordinates, is

$$\ddot\gamma^k + \sum_{ij} \Gamma_{ij}^k\,\dot\gamma^i\,\dot\gamma^j = 0.$$

(Note: $\ddot\gamma^k$ is the second derivative of $t \to \gamma^k(t)$, *not* the k^{th} component of $\ddot\gamma$.) The speed of such a curve γ is constant, and to stress this fact one can say that these are constant-speed geodesics, by opposition with general geodesics that can be reparametrized in an arbitrary way. Most of the time I shall just say "geodesics" for constant-speed geodesics. It is equivalent to say that a geodesic γ has constant speed, or that its length between two times $s < t$ is proportional to $t - s$.

We have just seen that minimizing curves have zero acceleration, and the converse is also true *locally*, that is if γ_1 is very close to γ_0. A curve which minimizes the action between its endpoints is called a *minimizing geodesic*, or minimal geodesic, or simply a geodesic. The Hopf-Rinow theorem guarantees that if the manifold M (seen as a metric space) is complete, then any two points in M are joined by at least one minimal geodesic. There might be several minimal geodesics joining two points x and y (to see this, consider two antipodal points on the sphere), but geodesics are

- *non-branching*: Two geodesics that are defined on a time interval $[0,t]$ and coincide on $[0,t']$ for some $t' > 0$ have to coincide on the whole of $[0,t]$. Actually, a stronger statement

holds true: The velocity of the geodesic at time $t = 0$ uniquely determines the final position at time $t = 1$ (this is a consequence of the uniqueness statement in the Cauchy-Lipschitz theorem);

- *locally unique*: For any given x, there is $r_x > 0$ such that any y in the ball $B_{r_x}(x)$ can be connected to x by a single geodesic $\gamma = \gamma^{x \to y}$, and then the map $y \longmapsto \dot\gamma(0)$ is a diffeomorphism (this corresponds to parametrize the endpoint by the initial velocity);

- *almost everywhere unique*: For any x, the set of points y that can be connected to x by several (minimizing!) geodesics is of zero measure. A way to see this is to note that the square distance function $d^2(x, \cdot)$ is locally semi-concave, and therefore differentiable almost everywhere. (See Chapter 10 for background about semi-concavity.)

The set $\Gamma_{x,y}$ of (minimizing, constant speed) geodesics joining x and y might not be single-valued, but in any case it is *compact* in $C([0,1], M)$, even if M is not compact. To see this, note that (i) the image of any element of $\Gamma_{x,y}$ lies entirely in the ball $B\big(x, d(x,y)\big)$, so $\Gamma_{x,y}$ is uniformly bounded, (ii) elements in $\Gamma_{x,y}$ are $d(x,y)$-Lipschitz, so they constitute an equi-Lipschitz family; (iii) $\Gamma_{x,y}$ is closed because it is defined by the equations $\gamma(0) = x$, $\gamma(1) = y$, $\mathcal{L}(\gamma) \le d(\gamma_0, \gamma_1)$ (the length functional $\mathcal{L}$ is not continuous with respect to uniform convergence, but it is lower semicontinuous, so an upper bound on the length defines a closed set); (iv) M is locally compact, so Ascoli's compactness theorem applies to functions with values in M.

A similar argument shows that for any two given compact sets K_s and K_t, the set of geodesics γ such that $\gamma_s \in K_s$ and $\gamma_t \in K_t$ is compact in $C([s,t]; M)$. So the Lagrangian action defined by $\mathcal{A}^{s,t}(\gamma) = \mathcal{L}(\gamma)^2/(t-s)$ is coercive in the sense of Definition 7.12.

Most of these statements can be generalized to the action coming from a Lagrangian function $L(x, v, t)$ on $TM \times [0,1]$, if L is C^2 and satisfies the classical conditions of Definition 7.6. In particular the associated cost functions will be continuous. Here is a sketch of proof: Let x and y be two given points, and let $x_k \to x$ and $y_k \to y$ be converging sequences. For any $\varepsilon > 0$, small enough,

$$c^{s,t}(x_k, y_k) \le c^{s,s+\varepsilon}(x_k, x) + c^{s+\varepsilon, t-\varepsilon}(x, y) + c^{t-\varepsilon, t}(y, y_k). \tag{7.32}$$

It is easy to show that there is a uniform bound K on the speeds of all minimizing curves which achieve the costs appearing above. Then the Lagrangian is uniformly bounded on these curves, so $c^{s,s+\varepsilon}(x_k, x) = O(\varepsilon)$, $c^{t-\varepsilon,t}(y, y_k) = O(\varepsilon)$. Also it does not affect much the Lagrangian (evaluated on candidate minimizers) to reparametrize $[s+\varepsilon, t-\varepsilon]$ into $[s,t]$ by a linear change of variables, so $c^{s+\varepsilon,t-\varepsilon}(x,y)$ converges to $c^{s,t}(x,y)$ as $s \to t$. This proves the upper semi-continuity, and therefore the continuity, of $c^{s,t}$.

In fact there is a finer statement: $c^{s,t}$ is superdifferentiable. This notion will be explained and developed later in Chapter 10.

After the Euclidean space, Riemannian manifolds constitute in some sense the most regular metric structure used by mathematicians. A Riemannian structure comes with many nice features (calculus, length, distance, geodesic equations); it also has a well-defined dimension n (the dimension of the manifold) and carries a natural volume, which can be defined equivalently as the n-dimensional Hausdorff measure associated with the distance, or by an explicit formula (in charts) involving the determinant of the metric.

Finsler structures constitute a generalization of the Riemannian structure: one has a differentiable manifold, with a norm on each tangent space T_xM, but that norm does not necessarily come from a scalar product. One can then define lengths of curves, the

induced distance as for a Riemannian manifold, and prove the existence of geodesics, but the geodesic equations are more complicated.

Another generalization is the notion of **length space** (or intrinsic length space), in which one does not necessarily have tangent spaces, yet one assumes the existence of a length $\mathcal{L}$ and a distance d which are compatible, in the sense that

$$\begin{cases} \mathcal{L}(\gamma) = \displaystyle\int_0^1 |\dot\gamma_t|\, dt, \qquad |\dot\gamma_t| := \limsup_{\varepsilon \to 0} \ \frac{d\big(\gamma_t, \gamma_{t+\varepsilon}\big)}{|\varepsilon|}, \\ \\ d(x,y) = \inf\Big\{ \mathcal{L}(\gamma); \quad \gamma_0 = x, \ \ \gamma_1 = y \Big\}. \end{cases}$$

In practice the following criterion is sometimes useful: A *complete* metric space $(\mathcal{X}, d)$ is a length space if and only if for any two points in $\mathcal{X}$, and any $\varepsilon > 0$ one can find an ε-midpoint of (x, y), that is a point m_ε such that

$$\left| \frac{d(x,y)}{2} - d(x, m_\varepsilon) \right| \le \varepsilon, \qquad \left| \frac{d(x,y)}{2} - d(y, m_\varepsilon) \right| \le \varepsilon.$$

Minimizing paths are fundamental objects in geometry. A length space in which any two points can be joined by a minimizing path, or **geodesic**, is called a strictly intrinsic length space, or geodesic space, or just (by abuse of language) a length space. There is a criterion in terms of midpoints: A complete metric space $(\mathcal{X}, d)$ is a geodesic space if and only if for any two points in $\mathcal{X}$ there is a midpoint, that is of course some $m \in X$ such that

$$d(x,m) = d(m,y) = \frac{d(x,y)}{2}.$$

There is another useful criterion: If the metric space $(\mathcal{X}, d)$ is a complete, *locally compact* length space, then it is geodesic. This is an analogue of the Hopf-Rinow theorem in Riemannian geometry. One can also reparametrize geodesic curves in such a way that their speed is constant, or equivalently that for all intermediate times s and t, their length between times s and t coincides with the distance between their positions at times s and t.

The same proof that I sketched for Riemannian manifolds applies in geodesic spaces, to show that the set $\Gamma_{x,y}$ of (minimizing, constant speed) geodesics joining x to y is compact; more generally, the set $\Gamma_{K_0 \to K_1}$ of geodesics γ with $\gamma_0 \in K_0$ and $\gamma_1 \in K_1$ is compact, as soon as K_0 and K_1 are compact. So there are important common points between the structure of a length space and the structure of a Riemannian manifold. From the practical point of view, some main differences are that (i) there is no available equation for geodesic curves, (ii) geodesics may "branch", (iii) there is no guarantee that geodesics between x and y are unique for y very close to x, (iv) there is neither a unique notion of dimension, nor a canonical reference measure, (v) there is no guarantee that geodesics will be almost everywhere unique.

Bibliographical Notes

There are plenty of classical textbooks on Riemannian geometry, with variable degree of pedagogy, among which the reader may consult [108], [142], [175]. For an introduction to the classical theory of calculus of variations in dimension 1, see for instance [160, Chapters 2-3], [83], or [113]. For an introduction to the Hamiltonian formalism in classical mechanics, one may use the very pedagogical treatise by Arnold [24], or the more complex

one by Thirring [351]. For an introduction to analysis in metric spaces, see Ambrosio and Tilli [18]. A wonderful introduction to the theory of length spaces can be found in Burago, Burago and Ivanov [81]. In the latter reference, a Riemannian manifold is defined as a length space which is locally isometric to $\mathbb{R}^n$ equipped with a quadratic form g_x depending smoothly on the point x. This definition is not standard, but it is equivalent to the classical definition, and in some sense more satisfactory if one wishes to emphasize the metric point of view.

I introduced the abstract concept of "coercive Lagrangian action" for the purpose of these notes, but this concept looks so natural to me that I would be somewhat surprised if it had not been previously discussed in the literature, maybe in a disguised form.

Probability measures on action-minimizing curves might look a bit scary when encountered for the first time, but they were actually rediscovered several times by various researchers, so they are arguably natural objects: See in particular the works by Bernot, Caselles and Morel [49] on irrigation problems; by Bangert [36] and Hohloch [211] on problems inspired by geometry and dynamical systems; by Ambrosio on transport equations with little or no regularity [12, 15]. In fact, in the context of partial differential equations, this approach already appears in the much earlier works of Brenier [71, 73, 74, 76] on the incompressible Euler equation and related systems. One technical difference is that Brenier considers probability measures on the huge (non-metrizable) space of measurable paths, while the other above-mentioned authors only consider much smaller spaces consisting of continuous, or Lipschitz-continuous functions. There are important subtleties with probability measures on non-metrizable spaces, and I strongly advise the reader to stay away from them.

Also in relation with the irrigation problem, Buttazzo, Santambrogio and Brancolini [84] have considered paths in the space of probability measures, however these authors do not really consider measures on trajectories.

The Hamilton–Jacobi equation with a quadratic cost function ($L(x, v, t) = |v|^2$) will be considered in more detail in Chapter 22; see in particular Proposition 22.13. For further information about Hamilton–Jacobi equations, there is an excellent book by Cannarsa and Sinestrari [94]. Of course the Hamilton–Jacobi equation is closely related to the concept of c-convexity: for instance, it is equivalent to say that ψ is c-convex, or that it is a solution at time 0 of the backward Hamilton–Jacobi semigroup starting at time 1 (with some arbitrary datum).

Measurable selection theorems provide conditions under which one may select elements satisfying certain conditions in a measurable way. The theorem which I used at the end of the proof of Proposition 7.15 is one of the most classical of these theorems; see Dellacherie [133] for a modern proof.

Interpolation arguments involving changes of variables have a long history. The concept and denomination of displacement interpolation was introduced by McCann [267] in the particular case of the quadratic cost in Euclidean space. Soon after, it was understood by Brenier that this procedure could formally be recast as an action minimization problem in the space of measures, which would reduce to the classical geodesic problem when the probability measures are Dirac masses. In Brenier's approach, the action is defined, at least formally, by

$$\mathbb{A}(\mu) = \inf_{v(t,x)} \left\{ \int_0^1 \int |v(t,x)|^2 \, d\mu_t(x) \, dt; \quad \frac{\partial \mu}{\partial t} + \nabla \cdot (v\mu) = 0 \right\},$$

and then one has the Benamou-Brenier formula

$$W_2(\mu_0, \mu_1)^2 = \inf \mathbb{A}(\mu),$$

where the infimum is taken among all paths $(\mu_t)_{0\le t\le 1}$ satisfying certain regularity conditions. Brenier himself gave two sketches of proof for this formula [42, 77], and another formal argument was suggested by Otto and the author [292, Section 3]. Rigorous proofs were later provided by several authors under various assumptions [365, Theorem 8.1] [201][15, Chapter 8] (the latter reference contains the most precise results). We shall come back to these formulas later on, after a more precise qualitative picture of optimal transport has emerged.

There was a rather amazing precursor to the idea of displacement interpolation, in the form of Nelson's theory of "stochastic mechanics". Nelson tried to build up a formalism in which quantum effects would be explained by stochastic fluctuations. For this purpose he considered an action minimization problem which was also studied by Guerra and Morato:

$$\inf\ \mathbb{E} \int_0^1 |\dot{X}_t|^2\, dt,$$

where the infimum is over all random paths $(X_t)_{0\le t\le 1}$ such that law $(X_0) = \mu_0$, law $(X_1) = \mu_1$, and in addition (X_t) solves the stochastic differential equation

$$\frac{dX_t}{dt} = \sigma\, \frac{dB_t}{dt} + \xi(t, X_t),$$

where $\sigma > 0$ is some coefficient, B_t is a standard Brownian motion, and ξ is a drift, which is an unknown in the problem. (So the minimization is over all possible couplings (X_0, X_1) but also over all drifts!) This formulation is very similar to the Benamou-Brenier formula just alluded to, only there is the additional Brownian noise in it, so it is more complex in some sense. Moreover, the expected value of the action is always infinite, so one has to renormalize it to make sense of Nelson's problem. Nelson made the incredible discovery that after a change of variables, minimizers of the action produced solutions of the free Schrödinger equation in $\mathbb{R}^n$. He developed his approach for some time, and finally gave up because it was introducing unpleasant nonlocal features. Part of the story can be found in Nelson's book [282], see also Carlen [95] and references therein.

It was Otto [290] who first explicitly reformulated the Benamou-Brenier formula as the equation for a geodesic distance on a Riemannian setting, from a formal point of view. Then Ambrosio and co-workers pointed out that if one is not interested in the equations of motion, but just in the geodesic property, it is simpler to use the metric notion of geodesic in a length space [15]. Those issues were also developed by other authors working with slightly different formalisms [15, 102, 97].

All the above-mentioned works were mainly concerned with displacement interpolation in $\mathbb{R}^n$. Agueh [1] also considered the case of cost $c(x, y) = |x - y|^p$ $(p > 1)$ in Euclidean space. Then displacement interpolation on Riemannian manifolds was studied, from a heuristic point of view, by Otto and myself [292]. Some useful technical tools were introduced in the field by Cordero-Erausquin, McCann and Schmuckenschläger [118] for Riemannian manifolds; Cordero-Erausquin adapted them to the case of rather general strictly convex cost functions in $\mathbb{R}^n$ [116].

The displacement interpolation for more general cost functions, arising from a smooth Lagrangian, was constructed by Bernard and Buffoni [47], who first introduced in this context Property (ii) in Theorem 7.19. At the same time, they made the explicit link with the Mather minimization problem, which will appear in the next chapters.

In all these works, displacement interpolation took place in a smooth structure, resulting in particular in the uniqueness (almost everywhere) of minimizing curves used in the interpolation. Displacement interpolation in length spaces, as presented in this chapter,

via the notion of dynamical transference plan, was developed more recently by Lott and myself [247]. Theorem 7.19 in these notes is new; it was essentially obtained by rewriting the proof in [247] with enough generality added to include the setting of Bernard and Buffoni.

The observation in Remark 7.25 came from a discussion with S. Evans, who pointed out to me that it was difficult, if not impossible, to get characterizations of random processes expressed in terms of the measures when working in state spaces that are not locally compact (such as the space of real trees). In spite of that remark, recently Lisini [244] was able to obtain representation theorems for general absolutely continuous paths $(\mu_t)_{0\le t\le 1}$ in the Wasserstein space $P_p(\mathcal{X})$ $(p > 1)$, as soon as $\int \|\dot\mu_t\|_{P_p}^p\, dt < \infty$, where $\mathcal{X}$ is just a Polish space and $\|\dot\mu_t\|_{P_p}$ is the metric speed in $P_p(\mathcal{X})$. He shows that such a curve may be written as $(e_t)_\# \mathit{\Pi}$, where $\mathit{\Pi}$ is the law of a random absolutely continuous curve γ; as a consequence, he could generalize Corollary 7.20 by removing the assumption of local compactness. Lisini also established a metric replacement for the relation of conservation of mass: For almost all t,

$$\mathbb{E}\,|\dot\gamma_t|^p \le \|\dot\mu\|_{P_p}^p.$$

He further applied his results to various problems about transport in infinite-dimensional Banach spaces.

Displacement interpolation in the case $p = 1$ is quite subtle because of the possibility of reparametrization; it was carefully discussed in the Euclidean space by Ambrosio [10]. Recently, Bernard and Buffoni [48] shed some new light on that issue by making explicit the link with the Mather–Mañé problem. Very roughly, the distance cost function is a typical representative of cost functions that arise from Lagrangians, if one also allows *minimization over the choice of the time-interval* $[0, T] \subset \mathbb{R}$ (rather than fixing, say, $T = 1$). This extra freedom accounts for the degeneracy of the problem.

8

The Monge–Mather shortening principle

Monge himself made the following important observation. Consider the transport cost $c(x, y) = |x - y|$ in the Euclidean plane, and two couples (x_1, y_1), (x_2, y_2), such that an optimal transport maps x_1 to y_1 and x_2 to y_2. (In our language, (x_1, y_1) and (x_2, y_2) are two points in the support of an optimal coupling π.) Then either all four points lie on a single line, or the two line segments $[x_1, y_1]$, $[x_2, y_2]$ do not cross, except maybe at their endpoints. The reason is easy to grasp: If the two lines would cross at a point which is not an endpoint of both lines, then, by triangular inequality we would have

$$|x_1 - y_2| + |x_2 - y_1| < |x_1 - y_1| + |x_2 - y_2|,$$

and this would contradict the fact that the support of π is c-cyclically monotone. Stated otherwise: Given two crossing line segments, we can *shorten* the total length of the paths by replacing these lines by the new transport lines $[x_1, y_2]$ and $[x_2, y_1]$.

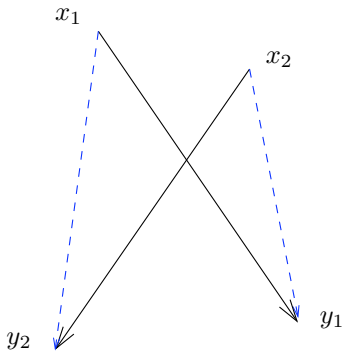


Fig. 8.1. Here the cost is Euclidean distance; if x_1 is sent to y_1 and x_2 to y_2, then it is cheaper to send x_1 to y_2 and x_2 to y_1.

Quadratic cost function

For cost functions that do not satisfy a triangular inequality, Monge's argument does not apply, and pathlines can cross. However, it is often the case that the crossing of the *curves* (with the time variable explicitly taken into account) is forbidden. Here is the most basic example: Consider the quadratic cost function in Euclidean space ($c(x, y) = |x - y|^2$), and let (x_1, y_1) and (x_2, y_2) belong to the support of some optimal coupling. By cyclical monotonicity,

$$|x_1 - y_1|^2 + |x_2 - y_2|^2 \le |x_1 - y_2|^2 + |x_2 - y_1|^2. \tag{8.1}$$

Let then

$$\gamma_1(t) = (1-t)\,x_1 + t\,y_1, \qquad \gamma_2(t) = (1-t)\,x_2 + t\,y_2$$

be the two minimizing curves respectively joining x_1 to y_1, and x_2 to y_2. Then it may happen that $\gamma_1(s) = \gamma_2(t)$ for some $s, t \in [0,1]$. But if there is a $t_0 \in (0,1)$ such that $\gamma_1(t_0) = \gamma_2(t_0) =: X$, then

$$\begin{aligned}
|x_1 - y_2|^2 + |x_2 - y_1|^2 &= |x_1 - X|^2 + |X - y_2|^2 - 2\,\langle X - x_1, X - y_2\rangle \\
&\qquad + |x_2 - X|^2 + |X - y_1|^2 - 2\langle X - x_2, X - y_1\rangle \\
&= [t_0^2 + (1-t_0)^2]\,\big(|x_1 - y_1|^2 + |x_2 - y_2|^2\big) + 4\,t_0(1-t_0)\,\langle x_1 - y_1, x_2 - y_2\rangle \\
&\le \big[t_0^2 + (1-t_0)^2 + 2\,t_0(1-t_0)\big]\,\big(|x_1 - y_1|^2 + |x_2 - y_2|^2\big) \\
&= |x_1 - y_1|^2 + |x_2 - y_2|^2,
\end{aligned}$$

and the inequality is strict unless $x_1 - y_1 = x_2 - y_2$, in which case $\gamma_1(t) = \gamma_2(t)$ for all t. But strict inequality contradicts (8.1). The conclusion is that *two distinct interpolation trajectories cannot meet at intermediate times.*

It is natural to ask whether this conclusion can be reinforced into a quantitative statement. The answer is yes. In fact there is a beautiful identity:

$$\begin{aligned}
\Big|\big((1-t)x_1 + t y_1\big) - \big((1-t)x_2 + t y_2\big)\Big|^2 &= (1-t)^2|x_1 - x_2|^2 + t^2|y_1 - y_2|^2 \\
&\quad + t(1-t)\Big(|x_1 - y_2|^2 + |x_2 - y_1|^2 - |x_1 - y_1|^2 - |x_2 - y_2|^2\Big).
\end{aligned} \tag{8.2}$$

To appreciate the consequences of (8.2), let

$$\gamma_1(t) = (1-t)x_1 + t y_1, \qquad \gamma_2(t) = (1-t)x_2 + t y_2.$$

Then (8.2) implies

$$\max\big(|x_1 - x_2|,\, |y_1 - y_2|\big) \le \max\left(\frac{1}{t},\, \frac{1}{1-t}\right)\,|\gamma_1(t) - \gamma_2(t)|.$$

Since $|\gamma_1(t) - \gamma_2(t)| \le \max(|x_1 - x_2|,\, |y_1 - y_2|)$ for all $t \in [0,1]$, one can conclude that

$$\forall t_0 \in (0,1), \qquad \sup_{0 \le t \le 1} |\gamma_1(t) - \gamma_2(t)| \le \max\left(\frac{1}{t_0},\, \frac{1}{1-t_0}\right)\,|\gamma_1(t_0) - \gamma_2(t_0)|. \tag{8.3}$$

(By the way, this inequality is easily seen to be optimal.) So the uniform distance between the *whole paths* γ_1 and γ_2 can be controlled by their distance at *some* time $t_0 \in (0,1)$.

General statement and applications to optimal transport

For the purpose of a seemingly different problem, Mather (not aware of Monge's work, neither of optimal transport) discovered an estimate which relies on the same idea as Monge's shortening argument — only much more complicated — for general cost functions on Lagrangian manifolds. He obtained a quantitative version of these estimates, in a form quite similar to (8.3).

Mather's proof uses three kinds of assumptions: (i) the existence of a second-order differential equation for minimizing curves; (ii) an assumption of regularity on the Lagrangian, and (iii) an assumption of strict convexity of the Lagrangian. To quantify the strict convexity, I shall use the following concept: A continuous function L on $\mathbb{R}^n$ will be said to be $(2+\kappa)$-*convex* if it satisfies a (strict) convexity inequality of the form

$$\frac{L(v)+L(w)}{2} - L\left(\frac{v+w}{2}\right) \geq K|v-w|^{2+\kappa}$$

for some constant $K > 0$.

The next statement is a slight generalization of Mather's estimate; if the reader finds it too dense, he or she can go directly to Corollary 8.2 which is simpler, and sufficient for the rest of this course.

Theorem 8.1 (Mather's shortening lemma). *Let M be a smooth Riemannian manifold, equipped with its geodesic distance d, and let $c(x,y)$ be a cost function on $M \times M$, defined by a Lagrangian $L(x,v,t)$ on $TM \times [0,1]$. Let x_1, x_2, y_1, y_2 be four points on M such that*

$$c(x_1,y_1) + c(x_2,y_2) \leq c(x_1,y_2) + c(x_2,y_1).$$

Let further γ_1 and γ_2 be two action-minimizing curves respectively joining x_1 to y_1 and x_2 to y_2. Let V be a bounded neighborhood of the graphs of γ_1 and γ_2 in $M \times [0,1]$, and S a strict upper bound on the maximal speed along these curves. Define

$$\mathcal{V} := \bigcup_{(x,t)\in V} \Big(x, B_S(0), t\Big) \subset TM \times [0,1].$$

In words, $\mathcal{V}$ is a neighborhood of γ_1 and γ_2, convex in the velocity variable.

Assume that (i) Minimizing curves for L are solution of a Lipschitz flow, in the sense of Definition 7.6 (e);

(ii) L is of class $C^{1,\alpha}$ in $\mathcal{V}$ with respect to the variables x and v, for some $\alpha \in (0,1]$ (so $\nabla_x L$ and $\nabla_v L$ are Hölder-α; Hölder-1 meaning Lipschitz);

(iii) L is $(2+\kappa)$-convex in $\mathcal{V}$, with respect to the v variable;

Then, for any $t_0 \in (0,1)$, there is a constant $C_{t_0} = C(L,\mathcal{V},t_0)$, and a positive exponent $\beta = \beta(\alpha,\kappa)$ such that

$$\sup_{0\leq t\leq 1} d\big(\gamma_1(t),\gamma_2(t)\big) \leq \; C_{t_0}\, d\big(\gamma_1(t_0),\gamma_2(t_0)\big)^{\beta}. \tag{8.4}$$

Furthermore, if $\alpha = 1$ and $\kappa = 0$, then $\beta = 1$ and $C_{t_0} = C(L,\mathcal{V})/\min(t_0,\, 1-t_0)$.

If L is of class C^2, superlinear and $\nabla_v^2 L > 0$ everywhere, then Assumption (iii) will be true by convexity with $\kappa = 0$, and, as we already discussed in Example 7.5, Assumption (i) will also be satisfied since minimizing curves will solve a differential equation of the form $\ddot\gamma(t) = f\Big(\gamma(t), \dot\gamma(t), t\Big)$, where f is Lipschitz in $\mathcal{V}$. So we have the following corollary:

Corollary 8.2 (Mather's shortening lemma again). *Let M be a smooth Riemannian manifold and let $L = L(x,v,t)$ be a C^2 Lagrangian on $TM \times [0,1]$, satisfying the classical assumptions of Definition 7.6, together with $\nabla_v^2 L > 0$. Let $c(x,y)$ be the cost function associated to L, and let $d(x,y)$ be the usual geodesic distance on M. Then, for any compact $K \subset M$ there is a constant C_K such that, whenever x_1, y_1, x_2, y_2 are four points in K with*

$$c(x_1, y_1) + c(x_2, y_2) \le c(x_1, y_2) + c(x_2, y_1),$$

and γ_1, γ_2 *are action-minimizing curves joining respectively* x_1 *to* y_1 *and* x_2 *to* y_2, *then for any* $t_0 \in (0,1)$,

$$\sup_{0\le t\le 1} d\big(\gamma_1(t), \gamma_2(t)\big) \le \frac{C_K}{\min(t_0,\, 1-t_0)}\, d\big(\gamma_1(t_0), \gamma_2(t_0)\big). \tag{8.5}$$

The short version of the conclusion is that the distance between γ_1 and γ_2 is controlled, *uniformly in* t, by the distance at *any* time $t_0 \in (0,1)$. In particular, the initial and final distance between these curves is controlled by their distance at any intermediate time. (But the final distance is *not* controlled by the initial distance!) Once again, inequalities (8.4) or (8.5) are quantitative versions of the qualitative statement that the two curves, if distinct, cannot cross except maybe at initial and final time.

Example 8.3. The cost function $c(x,y) = d(x,y)^2$ corresponds to the Lagrangian function $L(x,v,t) = |v|^2$, which obviously satisfies the assumptions of Corollary 8.2. In that case the exponent $\beta = 1$ is admissible. Moreover, it is natural to expect that the constant C_K can be controlled in terms of just a *lower bound on the sectional curvature* of M. I shall come back to this issue later in this chapter.

Example 8.4. The cost function $c(x,y) = d(x,y)^{1+\alpha}$ does not satisfy the assumptions of Corollary 8.2 for $0 < \alpha < 1$. Even if the associated Lagrangian $L(x,v,t) = |v|^{1+\alpha}$ is not smooth, the equation for minimizing curves is just the geodesic equation, so that Assumption (i) in Theorem 8.1 is still satisfied. Then, by tracking exponents in the proof of Theorem 8.1, one can find that (8.4) holds true with $\beta = (1+\alpha)/(3-\alpha)$. But this is far from optimal: By taking advantage of the homogeneity of the power function, one can prove that the exponent $\beta = 1$ is also admissible, for all $\alpha \in (0,1)$. (It is the constant, rather than the exponent, which deteriorates as $\alpha \downarrow 0$.) I shall explain this argument in the Appendix, in the Euclidean case, and leave the Riemannian case as a delicate exercise. This example suggests that Theorem 8.1 still leaves room for improvement.

The proof of Theorem 8.1 is a bit involved and before presenting it I prefer to discuss some applications in terms of optimal couplings.

Theorem 8.5 (The transport from intermediate times is Lipschitz). *On a Riemannian manifold* M, *let* c *be a cost function satisfying the assumptions of Corollary 8.2. Then, for any dynamical optimal transport* Π *supported in a compact set* K, *one has,* $\Pi \otimes \Pi(d\gamma\, d\widetilde{\gamma})$*-almost surely,*

$$\sup_{0\le t\le 1} d\big(\gamma(t), \widetilde{\gamma}(t)\big) \le C_K(t_0)\, d\big(\gamma(t_0), \widetilde{\gamma}(t_0)\big). \tag{8.6}$$

In particular, if $(\mu_t)_{0\le t\le 1}$ *is a displacement interpolation between any two compactly supported probability measures on* M, *and* $t_0 \in (0,1)$ *is given, then for any* $t \in [0,1]$ *the map*

$$T_{t_0\to t} :\ \gamma(t_0) \longmapsto \gamma(t)$$

is well-defined μ_{t_0}*-almost surely and Lipschitz continuous on its domain; and it is in fact the unique solution of the Monge problem between* μ_{t_0} *and* μ_t. *In other words, the coupling* $(\gamma(t_0), \gamma(t))$ *is an optimal deterministic coupling.*

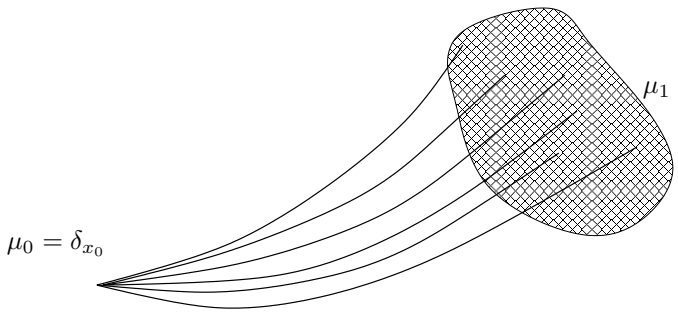


Fig. 8.2. On this example the map $\gamma(0) \to \gamma(1/2)$ is not well-defined, but the map $\gamma(1/2) \to \gamma(0)$ is well-defined and Lipschitz, just as the map $\gamma(1/2) \to \gamma(1)$. Also μ_0 is singular, but μ_t is absolutely continuous as soon as $t > 0$.

Example 8.6. On $\mathbb{R}^n$ with $c(x, y) = |x - y|^2$, let $\mu_0 = \delta_0$ and let $\mu = \text{law}\,(X)$ be arbitrary. Then it is easy to check that $\mu_t = \text{law}\,(tX)$, and in fact the random geodesic $\gamma(t)$ is just $t\gamma(1)$. So $\gamma(t) = t\gamma(t_0)/t_0$, which obviously provides a deterministic coupling.

Proof of Theorem 8.5. The proof consists only in formalizing things that by now may look essentially obvious to the reader. First, $(e_0, e_1, e_0, e_1)_\#(\Pi \otimes \Pi) = \pi \otimes \pi$, where π is an optimal coupling between μ_0 and μ_1. So if a certain property holds true $\pi \otimes \pi$-almost surely for quadruples, it also holds true $\Pi \otimes \Pi$-almost surely for the endpoints of couples of curves.

Since π is optimal, it is c-cyclically monotone (Theorem 5.9 (ii)), so, $\pi \otimes \pi(dx\, dy\, d\widetilde{x}\, d\widetilde{y})$-almost surely,

$$c(x, y) + c(\widetilde{x}, \widetilde{y}) \le c(x, \widetilde{y}) + c(\widetilde{x}, y).$$

Thus, $\Pi \otimes \Pi(d\gamma\, d\widetilde{\gamma})$-almost surely,

$$c(\gamma(0), \gamma(1)) + c(\widetilde{\gamma}(0), \widetilde{\gamma}(1)) \le c(\gamma(0), \widetilde{\gamma}(1)) + c(\widetilde{\gamma}(0), \gamma(1)).$$

Then (8.6) follows from Corollary 8.2.

Let S be the support of Π; by assumption this is a compact set. Since the inequality (8.6) defines a closed set of couples of geodesics, actually it has to hold true for *all* couples $(\gamma, \widetilde{\gamma}) \in S \times S$.

Now define the map $T_{t_0 \to t}$ on the compact set $e_{t_0}(S)$ (that is, the union of all $\gamma(t_0)$, when γ varies over the compact set S), by the formula $T_{t_0 \to t}(\gamma(t_0)) = \gamma(t)$. This map is well-defined, for if two geodesics γ and $\widetilde{\gamma}$ in the support of Π are such that $\gamma(t_0) = \widetilde{\gamma}(t_0)$, then inequality (8.6) imposes $\gamma = \widetilde{\gamma}$. The same inequality shows that $T_{t_0 \to t}$ is actually Lipschitz-continuous, with Lipschitz constant $C_K / \min(t_0, 1 - t_0)$.

All this shows that $(\gamma(t_0), T_{t_0}(\gamma(t_0)))$ is indeed a Monge coupling of (μ_{t_0}, μ_t), with a Lipschitz map. To complete the proof of the theorem, it only remains to check the uniqueness of the optimal coupling; but this follows from Theorem 7.27(iii). □

The second application is a result of "preservation of absolute continuity".

Theorem 8.7 (absolute continuity of displacement interpolation). *Let M be a Riemannian manifold, and let $L(x, v, t)$ be a C^2 Lagrangian on $TM \times [0, 1]$, bounded below, superlinear and strictly convex in the velocity variable, with $\nabla_v^2 L > 0$; let c be the associated cost function. Let μ_0 and μ_1 be two probability measures on M such that the optimal cost $C(\mu_0, \mu_1)$ is finite, and let $(\mu_t)_{0 \le t \le 1}$ be a displacement interpolation between μ_0 and μ_1. If*

either μ_0 or μ_1 is absolutely continuous with respect to the volume on M, then also μ_t is absolutely continuous for all $t \in (0,1)$.

Proof of Theorem 8.7. Let us assume for instance that μ_1 is absolutely continuous, and prove that μ_{t_0} is also absolutely continuous ($0 < t_0 < 1$).

First consider the case when μ_0 and μ_1 are compactly supported. Then the whole displacement interpolation is compactly supported, and Theorem 8.5 applies, so there is a Lipschitz map T such that $T_\# \mu_{t_0} = \mu_1$.

Now if N is a set of zero volume, the inclusion $N \subset T^{-1}(T(N))$ implies

$$\mu[N] \leq \mu\big[T^{-1}(T(N))\big] = (T_\# \mu)[T(N)] = \nu[T(N)], \tag{8.7}$$

and the latter quantity is 0 since $\mathrm{vol}\,[T(N)] \leq \|T\|_{\mathrm{Lip}}^n \,\mathrm{vol}\,[N] = 0$. Then $\nu[T(N)] = 0$ since ν is absolutely continuous with respect to vol. So (8.7) shows that $\mu[N] = 0$ for any Borel set N of zero volume, and this means precisely that μ is absolutely continuous.

Actually, the previous computation is not completely rigorous because $T(N)$ is not necessarily Borel measurable; but this is not serious since $T(N)$ can still be included in a negligible Borel set, and then the proof can be repaired in an obvious way.

Now let us turn to the general case where μ_0 and μ_1 are not assumed to be compactly supported. This situation will be handled by a restriction argument. Assume by contradiction that μ_{t_0} is not absolutely continuous. Then there exists a set Z_{t_0} with zero volume, such that $\mu_{t_0}[Z_{t_0}] > 0$. Let $\mathcal{Z} := \{\gamma \in \Gamma(M);\ \gamma_{t_0} \in Z_{t_0}\}$. Then

$$\Pi[\mathcal{Z}] = \mathbb{P}\,[\gamma_{t_0} \in Z_{t_0}] = \mu_{t_0}[Z_{t_0}] > 0.$$

By regularity, there exists a compact $\mathcal{K} \subset \mathcal{Z}$, such that $\Pi[\mathcal{K}] > 0$. Let then

$$\Pi' := \frac{1_{\mathcal{K}}\, \Pi}{\Pi[\mathcal{K}]},$$

and let $\pi' := (e_0, e_1)_\# \Pi'$ be the associated transference plan, and $\mu'_t = (e_t)_\# \Pi'$ the marginal of Π' at time t. In particular,

$$\mu'_1 \leq \frac{(e_0)_\# \Pi}{\Pi[\mathcal{K}]} = \frac{\mu_1}{\Pi[\mathcal{K}]},$$

so μ'_1 is still absolutely continuous.

Obviously, the support of Π' is included in that of Π, so Π' is concentrated on action-minimizing curves. It is easy to check that $(\mu'_t)_{0 \leq t \leq 1}$ is still a continuous path in the space of probability measures. On the other hand, by Theorem 4.5, π' is an optimal transference plan between μ'_0 and μ'_1. By Theorem 7.19, (μ'_t) is a displacement interpolation. Now, μ'_{t_0} is concentrated on $e_{t_0}(\mathcal{K}) \subset e_{t_0}(\mathcal{Z}) \subset Z_{t_0}$, so μ'_{t_0} is singular. But the first part of the proof rules out this possibility, because μ'_0 and μ'_1 are respectively supported in $e_0(\mathcal{K})$ and $e_1(\mathcal{K})$, which are compact, and μ'_1 is absolutely continuous. □

Proof of Mather's estimates

Now, let us turn to the proof of Theorem 8.1. It is certainly more important to grasp the idea of the proof than to follow the calculations, so the reader might at first reading be content with the following explanations and skip the rigorous proof.

Idea of the proof of Theorem 8.1. Assume, to fix the ideas, that γ_1 and γ_2 cross each other at a point m_0 and at time t_0. Close to m_0, these two curves look like two straight lines crossing each other, with respective velocities v_1 and v_2. Now cut these curves on the time interval $[t_0-\tau, t_0+\tau]$ and on that interval introduce "deviations" (like a plumber installing a new piece of pipe to short-cut a damaged region of a channel) that join the first curve to the second, and vice versa.

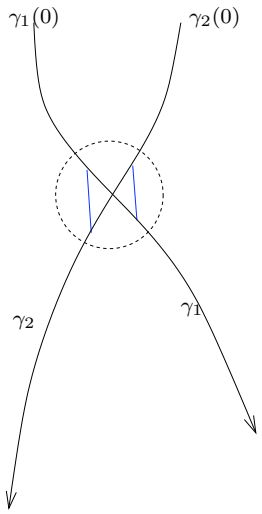


Fig. 8.3. Principle of Mather's proof: Let γ_1 and γ_2 be two action-minimizing curves. If at time t_0 the two curves γ_1 and γ_2 pass too close to each other, one can devise shortcuts (here drawn as straight lines).

This amounts to replacing (on a short interval of time) two curves with approximate velocities v_1 and v_2, by two curves with approximate velocities $(v_1 + v_2)/2$. Since the time-interval where the modification occurs is short, everything is concentrated in the neighborhood of (m_0, t_0), so the modification in the Lagrangian action of the two curves is approximately

$$(2\tau)\left(2\,L\left(m_0, \frac{v_1+v_2}{2}, t_0\right) - \Big[L(m_0, v_1, t_0) + L(m_0, v_2, t_0)\Big]\right).$$

Since $L(m_0, \cdot, t_0)$ is strictly convex, this quantity is negative if $v_1 \neq v_2$, which means that the total action has been *strictly improved* by the modification. But then $c(x_1, y_2) + c(x_2, y_1) < c(x_1, y_1) + c(x_2, y_2)$, in contradiction with our assumptions. The only possibility out is that $v_1 = v_2$, i.e. at the crossing point the curves have the same position and the same velocity; but then, since they are solutions of a second-order differential inequality, these curves have to coincide for all times. □

Now it only remains to make this argument quantitative: If the two curves pass close to each other at time t_0, then their velocities at that time will also be close to each other, and so the trajectories have to be coincide for all times in $[0, 1]$. Unfortunately this will not be so easy.

Rigorous proof of Theorem 8.1. **Step 1: Localization.** The goal of this step is to show that the problem reduces to a local computation, that can be performed as if we were in Euclidean space, and that it is sufficient to control the difference of the velocities at time t_0 (as in the above sketchy explanation). If the reader is ready to believe in these two statements, then he or she can go directly to Step 2.

For brevity, write $\gamma_1 \cup \gamma_2$ for the union of the images of the minimizing paths γ_1 and γ_2. For any point x in $\mathrm{proj}_M(V)$, there is a small ball $B_{r_x}(x)$ which is diffeomorphic to an open set in $\mathbb{R}^n$, and by compactness one can cover a neighborhood of $\gamma_1 \cup \gamma_2$ by a finite number of such balls B_j, each of them having radius no less than $\delta > 0$. Without loss of generality, all these balls are included in $\mathrm{proj}_M(V)$, and it can be assumed that whenever two points X_1 and X_2 in $\gamma_1 \cup \gamma_2$ are separated by a distance less than $\delta/4$, then there is one of the balls B_j that contains $B_{\delta/4}(X_1) \cup B_{\delta/4}(X_2)$.

Now, if $\gamma_1(t_0)$ and $\gamma_2(t_0)$ are separated by a distance at least $\delta/4$, then the conclusion is obvious. Otherwise, choose τ small enough that $\tau S \leq \delta/4$ (recall that S is the maximum speed along the curves); then on the time-interval $[t_0 - \tau, t_0 + \tau]$ the curves never leave the balls $B_{\delta/4}(X_1) \cup B_{\delta/4}(X_2)$, and therefore the whole trajectories of γ_1 and γ_2 on that time-interval have to stay within a single ball B_j. If one takes into account positions, velocities and time, the system is confined within $B_j \times B_S(0) \times [0,1] \subset \mathcal{V}$.

On any of these balls B_j, one can introduce a Euclidean system of coordinates, and perform all computations in that system (write L in those coordinates, etc.) The distance induced on B_j by that system of coordinates will not be the same as the original Riemannian distance, but it can be bounded from above and below by multiples thereof. So we can pretend that we are really working with a Euclidean metric, and all conclusions that are obtained, involving only what happens inside the ball B_j, will remain true up to changing the bounds. Then, for the sake of all computations, we can freely add points as if we were working in Euclidean space.

If it can be shown, in that system of coordinates, that

$$\big|\dot\gamma_1(t_0) - \dot\gamma_2(t_0)\big| \;\leq\; C\big|\gamma_1(t_0) - \gamma_2(t_0)\big|, \tag{8.8}$$

then this means that $(\gamma_1(t_0), \dot\gamma_1(t_0))$ and $(\gamma_2(t_0), \dot\gamma_2(t_0))$ are very close to each other in TM; more precisely they are separated by a distance which is $O\big(d(\gamma_1(t_0), \gamma_2(t_0))\big)$. Then by Assumption (i) and Cauchy-Lipschitz theory this bound will be propagated backward and forward in time, so the distance between $(\gamma_1(t), \dot\gamma_1(t))$ and $(\gamma_2(t), \dot\gamma_2(t))$ will remain bounded by $O\big(d(\gamma_1(t_0), \gamma_2(t_0))\big)$. So to conclude the argument it is sufficient to prove (8.8).

Step 2: Construction of shortcuts. First some notation: Let us write $x_1(t) = \gamma_1(t)$, $x_2(t) = \gamma_2(t)$, $v_1(t) = \dot\gamma_1(t)$, $v_2(t) = \dot\gamma_2(t)$, and also $X_1 = x_1(t_0)$, $V_1 = v_1(t_0)$, $X_2 = x_2(t_0)$, $V_2 = v_2(t_0)$. The goal is to control $|V_1 - V_2|$ by $|X_1 - X_2|$.

Let $x_{12}(t)$ be defined by

$$x_{12}(t) = \begin{cases} x_1(t) & \text{for } t \in [0, t_0 - \tau]; \\ \\ \frac{x_1(t)+x_2(t)}{2} + \left(\frac{\tau+t-t_0}{2\tau}\right)\left(\frac{x_2(t_0+\tau)-x_1(t_0+\tau)}{2}\right) + \left(\frac{\tau-t+t_0}{2\tau}\right)\left(\frac{x_1(t_0-\tau)-x_2(t_0-\tau)}{2}\right) \\ \qquad\qquad \text{for } t \in [t_0 - \tau, t_0 + \tau]; \\ x_2(t) & \text{for } t \in [t_0 + \tau, 1]. \end{cases}$$

Note that x_{12} is a continuous function of t; it is a path that starts along γ_1, then switches to γ_2. Let $v_{12}(t)$ stand for its time-derivative:

$$v_{12}(t) = \begin{cases} v_1(t) & \text{for } t \in [0, t_0 - \tau]; \\ \\ \frac{v_1(t)+v_2(t)}{2} + \frac{1}{2\tau}\left(\left[\frac{x_2(t_0-\tau)+x_2(t_0+\tau)}{2}\right] - \left[\frac{x_1(t_0-\tau)+x_1(t_0+\tau)}{2}\right]\right) \\ \qquad\qquad \text{for } t \in [t_0 - \tau, t_0 + \tau]; \\ v_2(t) & \text{for } t \in [t_0 + \tau, 1]. \end{cases}$$

Then the path $x_{21}(t)$ and its time-derivative $v_{21}(t)$ are defined symetrically. These definitions are rather natural: First we try to construct paths on $[t_0 - \tau, t_0 + \tau]$ whose velocity is about the half of the velocities of γ_1 and γ_2; then we correct these paths by adding simple functions (linear in time) to make them match the correct endpoints.

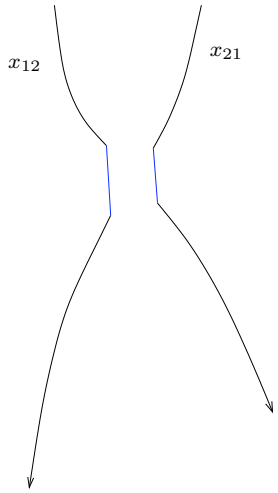


Fig. 8.4. The paths $x_{12}(t)$ and $x_{21}(t)$ obtained by using the shortcuts to switch from one original path to the other.

I shall conclude this step with some basic estimates about the paths x_{12} and x_{21} on the time-interval $[t_0 - \tau, t_0 + \tau]$. For a start, note that

$$x_{12} - \frac{x_1 + x_2}{2} = -\Big(x_{21} - \frac{x_1 + x_2}{2}\Big), \qquad v_{12} - \frac{v_1 + v_2}{2} = -\Big(v_{21} - \frac{v_1 + v_2}{2}\Big). \quad (8.9)$$

In the sequel, the symbol $O(m)$ will stand for any expression which is bounded by Cm, where C only depends on $\mathcal{V}$ and on the regularity bounds on the Lagrangian L on $\mathcal{V}$. From Cauchy-Lipschitz theory and Assumption (i),

$$|v_1 - v_2|(t) + |x_1 - x_2|(t) = O\Big(|X_1 - X_2| + |V_1 - V_2|\Big), \quad (8.10)$$

and then by plugging this back in the equation for minimizing curves we obtain

$$|\dot v_1 - \dot v_2|(t) = O\Big(|X_1 - X_2| + |V_1 - V_2|\Big).$$

Upon integration in times, these bounds imply

$$x_1(t) - x_2(t) = (X_1 - X_2) + O\big(\tau(|X_1 - X_2| + |V_1 - V_2|)\big); \quad (8.11)$$

$$v_1(t) - v_2(t) = (V_1 - V_2) + O\big(\tau(|X_1 - X_2| + |V_1 - V_2|)\big), \quad (8.12)$$

and therefore also

$$x_1(t) - x_2(t) = (X_1 - X_2) + (t - t_0)\,(V_1 - V_2) + O\big(\tau^2(|X_1 - X_2| + |V_1 - V_2|)\big). \quad (8.13)$$

As a consequence of (8.12), if τ is small enough (depending only on the Lagrangian L),

$$|v_1 - v_2|(t) \geq \frac{|V_1 - V_2|}{2} - O\big(\tau |X_1 - X_2|\big). \tag{8.14}$$

Next, from Cauchy-Lipschitz again,

$$x_2(t_0 + \tau) - x_1(t_0 + \tau) = X_2 - X_1 + \tau(V_2 - V_1) + O\big(\tau^2(|X_1 - X_2| + |V_1 - V_2|)\big);$$

and since a similar expression holds true with τ replaced by $-\tau$, one has

$$\left[\frac{x_1(t_0 + \tau) + x_2(t_0 + \tau)}{2}\right] - \left[\frac{x_1(t_0 - \tau) + x_2(t_0 - \tau)}{2}\right] = (X_2 - X_1) + O\Big(\tau^2(|X_1 - X_2| + |V_1 - V_2|)\Big). \tag{8.15}$$

It follows that

$$v_{12}(t) - \frac{v_1(t) + v_2(t)}{2} = O\Big(\frac{|X_1 - X_2|}{\tau} + \tau |V_1 - V_2|\Big). \tag{8.16}$$

After integration in time and use of (8.15) again, one obtains

$$\begin{aligned} x_{12}(t) - \frac{x_1(t) + x_2(t)}{2} &= \Big[x_{12}(t_0) - \frac{x_1(t_0) + x_2(t_0)}{2}\Big] + O\Big(|X_1 - X_2| + \tau^2 |V_1 - V_2|\Big) \\ &= O\Big(|X_1 - X_2| + \tau^2 |V_1 - V_2|\Big) \end{aligned} \tag{8.17}$$

In particular,

$$|x_{12} - x_{21}|(t) = O\Big(|X_1 - X_2| + \tau^2 |V_1 - V_2|\Big). \tag{8.18}$$

Step 3: Taylor formulas and regularity of L. Now I shall evaluate the behavior of L along the old and the new paths, using the regularity assumption (ii). From that point on, I shall drop the time variable for simplicity (but it is implicit in all the computations). First,

$$L(x_1, v_1) - L\left(\frac{x_1 + x_2}{2}, v_1\right) = \nabla_x L\left(\frac{x_1 + x_2}{2}, v_1\right) \cdot \left(\frac{x_1 - x_2}{2}\right) + O\big(|x_1 - x_2|^{1+\alpha}\big);$$

similarly

$$L(x_2, v_2) - L\left(\frac{x_1 + x_2}{2}, v_2\right) = \nabla_x L\left(\frac{x_1 + x_2}{2}, v_2\right) \cdot \left(\frac{x_1 - x_2}{2}\right) + O\big(|x_1 - x_2|^{1+\alpha}\big).$$

Moreover,

$$\nabla_x L\left(\frac{x_1 + x_2}{2}, v_1\right) - \nabla_x L\left(\frac{x_1 + x_2}{2}, v_2\right) = O(|v_1 - v_2|^\alpha).$$

The combination of these three identities, together with estimates (8.11) and (8.12), yields

$$\begin{aligned} \Big(L(x_1, v_1) + L(x_2, v_2)\Big) - \Big(L\Big(\frac{x_1 + x_2}{2}, v_1\Big) + L\Big(\frac{x_1 + x_2}{2}, v_2\Big)\Big) \\ &= O\Big(|x_1 - x_2|^{1+\alpha} + |x_1 - x_2|\,|v_1 - v_2|^\alpha\Big) \\ &= O\Big(|X_1 - X_2|^{1+\alpha} + \tau |V_1 - V_2|^{1+\alpha} + |X_1 - X_2|\,|V_1 - V_2|^\alpha \\ &\qquad + \tau^{1+\alpha}\,|V_1 - V_2|\,|X_1 - X_2|^\alpha\Big). \end{aligned}$$

Next, in an analogous way,

$$L(x_{12},v_{12})-L\Big(x_{12},\frac{v_1+v_2}{2}\Big)=\nabla_v L\Big(x_{12},\frac{v_1+v_2}{2}\Big)\cdot\big(v_{12}-\frac{v_1+v_2}{2}\big)+O\left(\big|v_{12}-\frac{v_1+v_2}{2}\big|^{1+\alpha}\right),$$

$$L(x_{21},v_{21})-L\Big(x_{21},\frac{v_1+v_2}{2}\Big)=\nabla_v L\Big(x_{21},\frac{v_1+v_2}{2}\Big)\cdot\big(v_{21}-\frac{v_1+v_2}{2}\big)+O\left(\big|v_{21}-\frac{v_1+v_2}{2}\big|^{1+\alpha}\right),$$

$$\nabla_v L\left(x_{12},\frac{v_1+v_2}{2}\right)-\nabla_v L\left(x_{21},\frac{v_1+v_2}{2}\right)=O\big(|x_{12}-x_{21}|^\alpha\big).$$

Combining this with (8.9), (8.16) and (8.18), one finds

$$\begin{aligned}\Big(L(x_{12},v_{12})+L(x_{21},v_{21})\Big)&-\Big(L\big(x_{12},\frac{v_1+v_2}{2}\big)+L\big(x_{21},\frac{v_1+v_2}{2}\big)\Big)\\&=O\Big(\big|v_{12}-\frac{v_1+v_2}{2}\big|^{1+\alpha}+\big|v_{12}-\frac{v_1+v_2}{2}\big|\,|x_{12}-x_{21}|^\alpha\Big)\\&=O\left(\frac{|X_1-X_2|^{1+\alpha}}{\tau^{1+\alpha}}+\tau^{1+\alpha}|V_1-V_2|^{1+\alpha}\right).\end{aligned}$$

After that,

$$\begin{aligned}L\Big(x_{12},\frac{v_1+v_2}{2}\Big)=L\Big(\frac{x_1+x_2}{2},\frac{v_1+v_2}{2}\Big)&+\nabla_x L\Big(\frac{x_1+x_2}{2},\frac{v_1+v_2}{2}\Big)\cdot\Big(x_{12}-\frac{x_1+x_2}{2}\Big)\\&+O\left(\big|x_{12}-\frac{x_1+x_2}{2}\big|^{1+\alpha}\right),\end{aligned}$$

$$\begin{aligned}L\Big(x_{21},\frac{v_1+v_2}{2}\Big)=L\Big(\frac{x_1+x_2}{2},\frac{v_1+v_2}{2}\Big)&+\nabla_x L\Big(\frac{x_1+x_2}{2},\frac{v_1+v_2}{2}\Big)\cdot\Big(x_{21}-\frac{x_1+x_2}{2}\Big)\\&+O\left(\big|x_{21}-\frac{x_1+x_2}{2}\big|^{1+\alpha}\right),\end{aligned}$$

and now by (8.9) the terms in ∇_x cancel each other exactly upon sommation, so the bound (8.17) leads to

$$\begin{aligned}\Big(L\big(x_{12},\frac{v_1+v_2}{2}\big)+L\big(x_{21},\frac{v_1+v_2}{2}\big)\Big)-2L\Big(\frac{x_1+x_2}{2},\frac{v_1+v_2}{2}\Big)&\\=O\left(\big|x_{21}-\frac{x_1+x_2}{2}\big|^{1+\alpha}\right)&\\=O\Big(|X_1-X_2|^{1+\alpha}+\tau^{2+2\alpha}|V_1-V_2|^{1+\alpha}\Big).&\end{aligned}$$

Step 4: Comparison of actions and strict convexity.
From our minimization assumption,

$$\mathcal{A}(x_1)+\mathcal{A}(x_2)\le\mathcal{A}(x_{12})+\mathcal{A}(x_{21}),$$

which of course can be rewritten

$$\begin{aligned}\int_{t_0-\tau}^{t_0+\tau}\Big(&L(x_1(t),v_1(t),t)+L(x_2(t),v_2(t),t)-L(x_{12}(t),v_{12}(t),t)\\&-L(x_{21}(t),v_{21}(t),t)\Big)\,dt\le 0.\end{aligned}\tag{8.19}$$

From Step 3, we can replace in the integrand all the positions by $(x_1 + x_2)/2$, and v_{12}, v_{21} by $(v_1 + v_2)/2$, up to a small error. Collecting the various error terms, and taking into account the smallness of τ, one obtains (dropping the t variable again)

$$\frac{1}{2\tau}\int_{t_0-\tau}^{t_0+\tau}\left\{L\Big(\frac{x_1+x_2}{2},v_1\Big)+L\Big(\frac{x_1+x_2}{2},v_2\Big)-2\,L\Big(\frac{x_1+x_2}{2},\frac{v_1+v_2}{2}\Big)\right\}dt \\ \leq C\Big(\frac{|X_1-X_2|^{1+\alpha}}{\tau^{1+\alpha}}+\tau|V_1-V_2|^{1+\alpha}\Big). \quad (8.20)$$

On the other hand, from the convexity condition (iii) and (8.14),

$$\frac{1}{2\tau}\int_{t_0-\tau}^{t_0+\tau}\left\{L\Big(\frac{x_1+x_2}{2},v_1\Big)+L\Big(\frac{x_1+x_2}{2},v_2\Big)-2\,L\Big(\frac{x_1+x_2}{2},\frac{x_1+x_2}{2}\Big)\right\}dt \quad (8.21)$$
$$\geq K\,\frac{1}{2\tau}\int_{t_0-\tau}^{t_0+\tau}|v_1-v_2|^{2+\kappa}\,dt$$
$$\geq K'\Big(|V_1-V_2|-A\tau|X_1-X_2|\Big)^{2+\kappa}.$$

If $|V_1 - V_2| \leq (A\tau + 1)|X_1 - X_2|$, then the proof is finished. Otherwise this means that $|V_1 - V_2| - A\tau|X_1 - X_2| \geq |V_1 - V_2|$, and then the combination of (8.19) and (8.21) implies

$$|V_1-V_2|^{2+\kappa}\leq C\Big(\frac{|X_1-X_2|^{1+\alpha}}{\tau^{1+\alpha}}+\tau|V_1-V_2|^{1+\alpha}\Big).$$

If $|V_1 - V_2| = 0$, then the proof is finished. Otherwise, the conclusion follows by choosing τ small enough that $C\tau|V_1 - V_2|^{1+\alpha} \leq (1/2)|V_1 - V_2|^{2+\kappa}$.

In the particular case when $\kappa = 0$ and $\alpha = 1$, one has

$$|V_1-V_2|^2\leq C\left(\frac{|X_1-X_2|^2}{\tau^2}+\tau|V_1-V_2|^2\right),$$

and if τ is small enough this implies just

$$|V_1-V_2|\leq C\,\frac{|X_1-X_2|}{\tau}. \quad (8.22)$$

The upper bound on τ depends on the regularity and strict convexity of τ in $\mathcal{V}$, but also on t_0 since τ cannot be greater than $\min(t_0, 1 - t_0)$. This is actually the only way in which t_0 explicitly enters the estimates. So inequality (8.22) concludes the argument. □

Remark 8.8. It is clear from the proof that one can compute a suitable exponent β, but also get more precise information upon the dependence of β on L by refining the smoothness assumptions, say $\nabla_x L$ is Hölder-α continuous in x, Hölder-δ continuous in v, etc.

Complement: Ruling out focalization by shortening

This section is about the application of the shortening technique to a classical problem in Riemannian geometry; it may be skipped at first reading.

Let M be a smooth Riemannian manifold and let $L = L(x, v, t)$ be a C^2 Lagrangian on $TM \times [0, 1]$, satisfying the classical assumptions of Definition 7.6, together with $\nabla_v^2 L > 0$. Let us assume that there is a well-defined map $(t, x_0, v_0) \to X_t(x_0, v_0)$, where $X_t(x_0, v_0)$ is

the position at time t of a system started from x_0 at time 0, with initial velocity v_0 and obeying the Euler–Lagrange equations of motion, given as a differential equation of second order with Lipschitz dependence.

It is said that there is **focalization** on another point $x' = X_{t'}(x_0, v_0)$, $t' > 0$, if the differential map $d_{v_0} X_{t'}(x_0, \cdot)$ is singular (not invertible). In words, this means that starting from x_0 it is very difficult to make the curve explore a whole neighborhood of x' by varying its initial velocity; instead, trajectories have a tendency to "concentrate" at time t' along certain preferred directions around x'.

The reader can test his or her understanding of the method exposed in the previous section by working out the details of the following

Problem 8.9 (Focalization is impossible before the cut locus). *With the same notation as before, let $\gamma : [0,1] \to M$ be a minimizing curve starting from some initial point x_0. By using the same strategy of proof as for Mather's estimates, show that, starting from x_0, focalization is impossible at $\gamma(t_*)$ if $0 < t_* < 1$. Hint: A possible reasoning is as follows:*

(a) Notice that the restriction of γ to $[0, t_*]$ is the unique minimizing curve on the time-interval $[0, t_*]$, joining x_0 to $x_* = \gamma(t_*)$;

(b) Take y close to x_* and introduce a minimizing curve $\widetilde{\gamma}$ on $[0, t_*]$, joining x_0 to y; show that the initial velocity $\widetilde{v}_0$ of $\widetilde{\gamma}$ is close to the initial velocity v_0 of γ if y is close enough to x_*;

(c) Bound the difference between the action of γ and the action of $\widetilde{\gamma}$ by $O(d(x_*, y))$; (recall that the size of the speeds along γ and $\widetilde{\gamma}$ is bounded by a uniform constant, depending only of the behavior of L in some compact set around γ)

(d) Construct a path $x_0 \to \gamma(1)$ by first going along $\widetilde{\gamma}$ up to time $t = t_* - \tau$ (τ small enough), then using a shortcut from $\widetilde{\gamma}(t_* - \tau)$ to $\gamma(t_* + \tau)$, finally going along γ up to time 1. Show that the gain of action is at least of the order of $\tau |V - \widetilde{V}|^2 - O(d(x_*, y)^2/\tau)$, where $V = \dot{\gamma}(t_*)$ and $\widetilde{V} = \dot{\widetilde{\gamma}}(t_*)$. Deduce that $|V - \widetilde{V}|$ is at most of the order $O(d(x_*, y)/\tau)$.

(e) Conclude that $|v_0 - \widetilde{v_0}| = O(d(x_*, y)/\tau)$. Use a contradiction argument to deduce that the differential map $d_{v_0} X_t(x_0, \cdot)$ is invertible, and more precisely that its inverse is of size $O((1 - t_*)^{-1})$ as a function of t_*.

In the important case when $L(x, v, t) = |v|^2$, what we have proven is a well-known result in Riemannian geometry; to explain it I shall first recall the notions of **cut locus** and **focal points**.

Let γ be a minimizing geodesic, and let t_c be the largest time such that for all $t < t_c$, γ is minimizing between γ_0 and γ_t. Roughly speaking, $\gamma(t_c)$ is the first point along the geodesic ceases to be minimizing; γ may or may not be minimizing between $\gamma(0)$ and $\gamma(t_c)$, but it is certainly not minimizing between $\gamma(0)$ and $\gamma(t_c + \varepsilon)$, for any $\varepsilon > 0$. Then the point $\gamma(t_c)$ is said to be a cut point of γ_0 along γ. When the initial position x_0 of the geodesic is fixed and the geodesic varies, the set of all cut points constitutes the *cut locus* of x_0.

Next, two points x_0 and x' are said to be *focal* (or conjugate) if x' can be written as $\exp_{x_0}(t' v_0)$, where the differential $d_{v_0} \exp_{x_0}(t' \cdot)$ is *not invertible*. As before, this means that x' can be obtained from x_0 by a geodesic γ with $\dot{\gamma}(0) = v_0$, such that it is difficult to explore a whole neighborhood of x' by slightly changing the initial velocity v_0.

With these notions, the main result of Problem 8.9 can be summarized as follows: *Focalization never occurs before the cut locus.* It can occur either at the cut locus, or after.

Example 8.10. Consider the sphere S^2. Let N be the North Pole, then it has only one cut point, which is also its focal point, namely the South Pole S. Now fix a geodesic γ going from $\gamma(0) = N$ to $\gamma(1) = S$, and deform your sphere out of a neighborhood of $\gamma[0,1]$, so as to dig a shortcut that allows to go from N to $\gamma(1/2)$ in a more efficient way than using γ. This will create a new cut point along γ, and S will not be a cut point along γ any longer (it might still be a cut point along some other geodesic). On the other hand, S will still be the only focal point along γ.

Remark 8.11. If x and y are not conjugate, and joined by a unique minimizing geodesic γ, then it is easy to show that there is a neighborhood U of y such that any z in U is also joined to x by a unique minimizing geodesic. Indeed, any minimizing geodesic has to be close to γ, therefore its initial velocity should be close to $\dot\gamma_0$; and by the local inversion theorem, there are neighborhoods W_0 of $\dot\gamma_0$ and U of y such that there is a unique correspondence between the initial velocity $\dot\gamma \in W_0$ of a minimizing curve starting from x, and the final point $\gamma(1) \in U$. This shows that the cut locus of a point x can be separated into two categories:

(a) those points y for which there are at least two distinct minimizing geodesics going from x to y;

(b) those points y for which there is a unique minimizing geodesic, but which are focal points of x.

Introduction to Mather's theory

In this section I shall present an application of Theorem 8.1 to the theory of Lagrangian dynamical systems. This is mainly to give the reader an idea of Mather's motivations, and to let him or her better understand the link between optimal transport and Mather's theory. These results will not play any role in the sequel of the notes.

Theorem 8.12 (Lipschitz graph Theorem). *Let M be a compact Riemannian manifold, let $L = L(x,v,t)$ be a Lagrangian function on $TM \times \mathbb{R}$, and let $T > 0$, such that*

(a) L is T-periodic in the t variable, i.e. $L(x,v,t+T) = L(x,v,t)$;

(b) L is of class C^2 in all variables;

(c) $\nabla_v^2 L$ is strictly positive everywhere, and L is superlinear in v.

Define as usual the action by $\mathcal{A}^{s,t}(\gamma) = \int_s^t L(\gamma_\tau, \dot\gamma_\tau, \tau)\, d\tau$. Let $c^{s,t}$ be the associated cost function on $M \times M$, and $C^{s,t}$ the corresponding optimal cost functional on $P(M) \times P(M)$.

Let $\overline{\mu}$ be a probability measure solving the minimization problem

$$\inf_{\mu \in P(\mathcal{X})} C^{0,T}(\mu, \mu), \tag{8.23}$$

and let $(\mu_t)_{0 \le t \le T}$ be a displacement interpolation between $\mu_0 = \overline{\mu}$ and $\mu_T = \overline{\mu}$. Extend (μ_t) into a T-periodic curve $\mathbb{R} \to P(M)$ defined for all times. Then

(i) For all $t_0 < t_1$, $(\mu_t)_{t_0 \le t \le t_1}$ still defines a displacement interpolation;

(ii) The optimal transport cost $C^{t,t+T}(\mu_t, \mu_t)$ is independent of t;

(iii) For any $t_0 \in \mathbb{R}$, and for any $k \in \mathbb{N}$, μ_{t_0} is a minimizer for $C^{t_0, t_0+kT}(\mu, \mu)$.

Moreover, there is a random curve $(\gamma_t)_{t \in \mathbb{R}}$, such that

(iv) For all $t \in \mathbb{R}$, $\text{law}\,(\gamma_t) = \mu_t$;

(v) For all times $t_0 < t_1$, the curve $(\gamma_t)_{t_0 \le t \le t_1}$ is action-minimizing;

(vi) The map $\gamma_0 \to \dot\gamma_0$ is well-defined and Lipschitz;

Remark 8.13. Since $c^{0,T}$ is not assumed to be nonnegative, the optimal transport problem (8.23) is not trivial.

Remark 8.14. If L does not depend on t, then one can apply the previous result for any $T = 2^{-\ell}$, and then use a compactness argument to construct a constant curve $(\mu_t)_{t\in\mathbb{R}}$ satisfying Properties (i)-(vi) above. In particular μ_0 is a *stationary measure* for the Lagrangian system.

Before giving its proof, let me explain briefly why Theorem 8.12 is interesting from the point of view of the dynamics. A trajectory of the dynamical system defined by the Lagrangian L is a curve γ which is locally action-minimizing; that is, one can cover the time-interval by small subintervals on which the curve is action-minimizing. It is a classical problem in mechanics to construct and study periodic trajectories having certain given properties. Theorem 8.12 does not construct a periodic trajectory, but at least it constructs a random trajectory γ (or equivalently a probability measure $\mathit{\Pi}$ on the set of trajectories) which is periodic *on the mean*: The law μ_t of γ_t satisfies $\mu_{t+T} = \mu_t$. This can also be thought of as a probability measure $\mathit{\Pi}$ on the set of all possible trajectories of the system.

Of course this in itself is not too striking, since there may be a great deal of invariant measures for a dynamical system, and some of them are often easy to construct. The important point in the conclusion of Theorem 8.12 is that the curve γ is not "too random", in the sense that the random variable $(\gamma(0), \dot\gamma(0))$ takes values in a Lipschitz *graph*. (If $(\gamma(0), \dot\gamma(0))$ were a deterministic element in TM, this would mean that $\mathit{\Pi}$ just sees a single periodic curve. Here we may have an infinite collection of curves, but still it is not "too large".)

Another remarkable property of the curves γ is the fact that the minimization property holds along *any* time-interval in $\mathbb{R}$, not necessarily small.

Example 8.15. Let M be a compact Riemannian manifold, and $L(x, v, t) = |v|^2/2 - V(x)$, where V has a unique maximum x_0. Then Mather's procedure selects the probability measure δ_{x_0}, and the stationary curve $\gamma \equiv x_0$.

It is a natural question whether we can construct more "interesting" measures and curves by Mather's procedure. A way to do so is to change the Lagrangian, for instance by replacing $L(x, v, t)$ by $L_\omega := L(x, v, t) + \omega(x) \cdot v$, where ω is a vector field on M. Indeed,

- If ω is closed (as a differential form), that is if $\nabla\omega$ is a symmetric operator, then L_ω and L have the same Euler–Lagrange equations, so the associated dynamical system is the same;

- If ω is exact, that is if $\omega = \nabla f$ for some function $f : M \to \mathbb{R}$, then L_ω and L have the same minimizing curves.

As a consequence, one may explore various parts of the dynamics by letting ω vary over the finite-dimensional group obtained by taking the quotient of the closed forms by the exact forms. In particular, one can make sure that the expected mean "rotation number" $\mathbb{E}\,\frac{1}{T}\int_0^T \dot\gamma\, dt$ takes nontrivial values as ω varies.

Proof of Theorem 8.12. First, $C^{0,T}(\mu, \mu)$ is clearly a lower semi-continuous function of μ, and it is bounded below by $T(\inf L) > -\infty$, so the minimization problem (8.23) does admit a solution.

So define $\mu_0 = \mu_T = \overline{\mu}$, then define μ_t by displacement interpolation for $0 < t < T$, and extend the result by periodicity.

Let $k \in \mathbb{N}$ be given and let $\widetilde{\mu}$ be a minimizer for the minimization problem

$$\inf_{\mu \in P(M)} C^{0,kT}(\mu, \mu).$$

We shall see later that actually $\overline{\mu}$ is a solution of this problem. For the moment, let $(\widetilde{\mu}_t)_{t\in\mathbb{R}}$ be obtained first by taking a displacement interpolation between $\widetilde{\mu}_0 = \widetilde{\mu}$ and $\widetilde{\mu}_{kT} = \widetilde{\mu}$; and then by extending the result by kT-periodicity.

On one hand,

$$C^{0,kT}(\overline{\mu}, \overline{\mu}) = C^{0,kT}(\mu_0, \mu_{kT}) \leq \sum_{j=0}^{k-1} C^{jT,(j+1)T}(\mu_{jT}, \mu_{(j+1)T}) = k\, C^{0,1}(\overline{\mu}, \overline{\mu}). \tag{8.24}$$

On the other hand, by definition of $\overline{\mu}$,

$$C^{0,T}(\mu, \mu) \leq C^{0,T}\Big(\frac{1}{k}\sum_{j=0}^{k-1} \widetilde{\mu}_{jT},\ \frac{1}{k}\sum_{j=0}^{k-1} \widetilde{\mu}_{jT}\Big) = C^{0,T}\Big(\frac{1}{k}\sum_{j=0}^{k-1} \widetilde{\mu}_{jT},\ \frac{1}{k}\sum_{j=0}^{k-1} \widetilde{\mu}_{(j+1)T}\Big). \tag{8.25}$$

Since $C^{0,T}(\mu, \nu)$ is a convex function of (μ, ν) (Theorem 4.7),

$$\begin{aligned} C^{0,T}\Big(\frac{1}{k}\sum_{j=0}^{k-1} \widetilde{\mu}_{jT},\ \frac{1}{k}\sum_{j=0}^{k-1} \widetilde{\mu}_{(j+1)T}\Big) &\leq \frac{1}{k}\sum_{j=0}^{k-1} C^{jT,(j+1)T}(\widetilde{\mu}_{jT}, \widetilde{\mu}_{(j+1)T}) \\ &= \frac{1}{k} C^{0,kT}(\widetilde{\mu}_0, \widetilde{\mu}_{kT}), \end{aligned} \tag{8.26}$$

where the last equality is a consequence of Property (ii) in Theorem 7.19. Inequalities (8.25) and (8.26) together imply

$$C^{0,1}(\overline{\mu}, \overline{\mu}) \leq \frac{1}{k}\, C^{0,kT}(\widetilde{\mu}_0, \widetilde{\mu}_{kT}) = \frac{1}{k}\, C^{0,kT}(\widetilde{\mu}, \widetilde{\mu}).$$

Since the reverse inequality holds true by (8.24), in fact all the inequalities in (8.24), (8.25) and (8.26) have to be equalities. In particular,

$$C^{0,kT}(\mu_0, \mu_{kT}) = \sum_{j=0}^{k-1} C^{jT,(j+1)T}(\mu_{jT}, \mu_{(j+1)T}). \tag{8.27}$$

Let us now check that the identity

$$C^{t_1,t_2}(\mu_{t_1}, \mu_{t_2}) + C^{t_2,t_3}(\mu_{t_2}, \mu_{t_3}) = C^{t_1,t_3}(\mu_{t_1}, \mu_{t_3}) \tag{8.28}$$

holds true for any three intermediate times $t_1 < t_2 < t_3$. By periodicity, it suffices to do this for $t_1 \geq 0$. If $0 \leq t_1 < t_2 < t_3 \leq T$, then (8.28) is true by the property of displacement interpolation. If $jT \leq t_1 < t_2 < t_3 \leq (j+1)T$, this is also true because of the T-periodicity. In the remaining cases, we may choose k large enough that $t_3 \leq kT$. Then

$$\begin{aligned} C^{0,kT}(\mu_0, \mu_{kT}) &\leq C^{0,t_1}(\mu_0, \mu_{t_1}) + C^{t_1,t_3}(\mu_{t_1}, \mu_{t_3}) + C^{t_3,kT}(\mu_{t_3}, \mu_{kT}) \\ &\leq C^{0,t_1}(\mu_0, \mu_{t_1}) + C^{t_1,t_2}(\mu_{t_1}, \mu_{t_2}) + C^{t_2,t_3}(\mu_{t_2}, \mu_{t_3}) + C^{t_3,kT}(\mu_{t_3}, \mu_{kT}) \\ &\leq \sum C^{s_j,s_{j+1}}(\mu_{s_j}, \mu_{s_{j+1}}), \end{aligned} \tag{8.29}$$

where the times s_j are obtained by ordering of $\{0, T, 2T, \dots, kT\} \cup \{t_1, t_2, t_3\}$. On each time-interval $[\ell T, (\ell+1)T]$ we know that (μ_t) is a displacement interpolation, so we can apply Theorem 7.19(ii), and as a result bound the right-hand side of (8.29) by

$$\sum_{\ell} C^{\ell T, (\ell+1)T}(\mu_{\ell T}, \mu_{(\ell+1)T}). \tag{8.30}$$

(Consider for instance the particular case when $0 < t_1 < t_2 < T < t_3 < 2T$; then one can write $C^{0,t_1} + C^{t_1,t_2} + C^{t_2,T} = C^{0,T}$, and also $C^{T,t_3} + C^{t_3,2T} = C^{T,2T}$. So $C^{0,t_1} + C^{t_1,t_2} + C^{t_2,T} + C^{T,t_3} + C^{t_3,2T} = C^{0,T} + C^{T,2T}$.)

But (8.30) is just $C^{0,kT}(\mu_0, \mu_{kT})$, as shown in (8.27). So there is in fact equality in all these inequalities, and (8.28) follows. Then by Theorem 7.19, (μ_t) defines a displacement interpolation between any two of its intermediate values. This proves (i). At this stage we have also proven (iii) in the case when $t_0 = 0$.

Now for any $t \in \mathbb{R}$, one has, by (8.28) and the T-periodicity,

$$\begin{aligned} C^{0,T}(\mu_0, \mu_T) = C^{0,t}(\mu_0, \mu_t) + C^{t,T}(\mu_t, \mu_T) = C^{t,T}(\mu_t, \mu_T) + C^{T,t+T}(\mu_T, \mu_{t+T}) \\ = C^{t,t+T}(\mu_t, \mu_{t+T}), \end{aligned}$$

which proves (ii).

Next, let t_0 be given, and repeat the same whole procedure with the initial time 0 replaced by t_0: That is, introduce a minimizer $\widetilde{\mu}$ for $C^{t_0,t_0+T}(\mu, \mu)$, etc. This gives a curve $(\widetilde{\mu}_t)_{t\in\mathbb{R}}$ with the property that $C^{t,t+T}(\widetilde{\mu}_t, \widetilde{\mu}_t) = C^{0,T}(\widetilde{\mu}_0, \widetilde{\mu}_0)$. It follows that

$$\begin{aligned} C^{t_0,t_0+T}(\mu_{t_0}, \mu_{t_0+T}) = C^{0,T}(\overline{\mu}, \overline{\mu}) \le C^{0,T}(\widetilde{\mu}_0, \widetilde{\mu}_0) \\ = C^{t_0,t_0+T}(\widetilde{\mu}_{t_0}, \widetilde{\mu}_{t_0}) = C^{t_0,t_0+T}(\widetilde{\mu}, \widetilde{\mu}) \le C^{t_0,t_0+T}(\mu_{t_0}, \mu_{t_0}). \end{aligned}$$

So there is equality everywhere, and μ_{t_0} is indeed a minimizer for $C^{t_0,t_0+T}(\mu, \mu)$. This proves the remaining part of (iii).

Next, let $(\gamma_t)_{0\le t\le T}$ be a random minimizing curve on $[0, T]$, such that $\mathrm{law}\,(\gamma_t) = \mu_t$, as in Theorem 7.19. For each k, define $(\gamma^k_t)_{kT\le t\le (k+1)T}$ as a copy of $(\gamma_t)_{0\le t\le T}$. Since μ_t is T-periodic, $\mathrm{law}\,(\gamma^k_{kT}) = \mathrm{law}\,(\gamma^k_{(k+1)T}) = \mu_0$, for all k. So we can *glue together* these random curves, just as in the proof of Theorem 7.27, and get random curves $(\gamma_t)_{t\in\mathbb{R}}$ such that $\mathrm{law}\,(\gamma_t) = \mu_t$ for all $t \in \mathbb{R}$, and each curve $(\gamma_t)_{kT\le t\le (k+1)T}$ is action-minimizing. Property (iv) is then satisfied by construction.

Property (v) can be proven by a principle which we already used in the proof of Theorem 7.19. Let us check for instance that γ is minimizing on $[0, 2T]$. For this one has to show that (almost surely)

$$c^{t_1,t_2}(\gamma_{t_1}, \gamma_{t_2}) + c^{t_2,t_3}(\gamma_{t_2}, \gamma_{t_3}) = c^{t_1,t_3}(\gamma_{t_1}, \gamma_{t_3}), \tag{8.31}$$

for any choice of intermediate times $t_1 < t_2 < t_3$ in $[0, 2T]$. Assume, without real loss of generality, that $0 < t_1 < t_2 < T < t_3 < 2T$. Then

$$\begin{aligned} C^{t_1,t_3}(\mu_{t_1}, \mu_{t_3}) &\le \mathbb{E}\, c^{t_1,t_3}(\gamma_{t_1}, \gamma_{t_3}) \\ &\le \mathbb{E}\left[c^{t_1,t_2}(\gamma_{t_1}, \gamma_{t_2}) + c^{t_2,t_3}(\gamma_{t_2}, \gamma_{t_3})\right] \\ &\le \mathbb{E}\, c^{t_1,t_2}(\gamma_{t_1}, \gamma_{t_2}) + \mathbb{E}\, c^{t_2,T}(\gamma_{t_2}, \gamma_T) + \mathbb{E}\, c^{T,t_3}(\gamma_T, \gamma_{t_3}) \\ &= C^{t_1,t_2}(\mu_{t_1}, \mu_{t_2}) + C^{t_1,t_2}(\mu_{t_1}, \mu_{t_2}) + C^{t_2,T}(\mu_{t_2}, \mu_T) + C^{T,t_3}(\mu_T, \mu_{t_3}) \\ &= C^{t_1,t_3}(\mu_{t_1}, \mu_{t_3}), \end{aligned}$$

where the property of optimality of the path $(\mu_t)_{t\in\mathbb{R}}$ was used in the last step. So all these inequalities are equalities, and in particular

$$\mathbb{E}\left[c^{t_1,t_3}(\gamma_{t_1},\gamma_{t_3}) - c^{t_1,t_2}(\gamma_{t_1},\gamma_{t_2}) - c^{t_2,t_3}(\gamma_{t_2},\gamma_{t_3})\right] = 0.$$

Since the integrand is nonpositive, it has to vanish almost surely. So (8.31) is satisfied almost surely, for given t_1, t_2, t_3. Then the same inequality holds true almost surely for all choices of rational times t_1, t_2, t_3; so by continuity of γ it holds true almost surely for all times. This concludes the proof of (v).

From general principles of Lagrangian mechanics, there is a uniform bound on the speeds of all the curves $(\gamma_t)_{-T\le t\le T}$ (this is because γ_{-T} and γ_T lie in a compact set). So for any given $\varepsilon > 0$ we can find δ such that $0 \le t \le \delta$ implies $d(\gamma_0,\gamma_t) \le \varepsilon$. Then if ε is small enough the map $(\gamma_0,\gamma_t) \to (\gamma_0,\dot\gamma(0))$ is Lipschitz. (This is another well-known fact in Lagrangian mechanics.) But from Theorem 8.5, applied with the intermediate time $t_0 = 0$ on the time-interval $[-T,T]$, we know that $\gamma_0 \longmapsto \gamma_t$ is well-defined (almost surely) and Lipschitz continuous. It follows that $\gamma_0 \to \dot\gamma_0$ is also Lipschitz continuous. This concludes the proof of Theorem 8.12. □

The story does not end up here. First, there is a powerful *dual point of view* to Mather's theory, based on solutions to the dual Kantorovich problem; this is a maximization problem defined by

$$\sup \int (\phi - \psi)\, d\mu, \tag{8.32}$$

where the supremum is over all probability measures μ on M, and all pairs of Lipschitz functions (ψ,ϕ) such that

$$\forall (x,y) \in M\times M, \qquad \phi(y) - \psi(x) \le c^{0,T}(x,y).$$

Next, Theorem 8.12 suggests that some objects related to optimal transport might be interesting to describe a Lagrangian system. This is indeed the case, and the notions defined below are useful and well-known in the theory of dynamical systems:

Definition 8.16 (Useful transport quantities describing a Lagrangian system). *For each displacement interpolation $(\overline{\mu}_t)_{t\ge0}$ as in Theorem 8.12, define*

(i) the Mather critical value as the opposite of the mean optimal transport cost:

$$-M = \overline{c} := \frac{1}{T} C^{0,T}(\overline{\mu},\overline{\mu}) = \frac{1}{kT} C^{0,kT}(\overline{\mu},\overline{\mu}); \tag{8.33}$$

(ii) the Mather set as the closure of the union of the supports of all measures $V_\#\overline{\mu}_0$, where $(\overline{\mu}_t)_{t\ge0}$ is a displacement interpolation as in Theorem 8.12 and V is the Lipschitz map $\gamma_0 \to (\gamma_0,\dot\gamma_0)$;

(iii) the Aubry set as the set of all $(\gamma_0,\dot\gamma_0)$ such that there exists a solution (ϕ,ψ) of the dual problem (8.32) *such that $H_+^{0,T}\psi(\gamma_1) - \psi(\gamma_0) = c^{0,T}(\gamma_0,\gamma_1)$.*

Up to the change of variables $(\gamma_0,\dot\gamma_0) \to (\gamma_0,\gamma_1)$, the Mather and Aubry sets are just the same as $\Gamma_{\min}$ and $\Gamma_{\max}$ appearing in the bibliographical notes of Chapter 5.

Example 8.17. Take a one-dimensional pendulum. For small values of the total energy, the pendulum is confined in a periodic motion, making just small oscillations, going back and forth around its equilibrium position and describing an arc of circle in physical space. For large values, it also has a periodic motion but now it goes always in the same direction,

and describes a complete circle ("revolution") in physical space. But if the system is given just the right amount of energy, it will describe a trajectory that is intermediate between these two regimes, and consists in going from the vertical upward position (at time $-\infty$) to the vertical upward position again (at time $+\infty$) after exploring all intermediate angles. There are two such trajectories (one clockwise, and one counterclockwise), which can be called revolutions of infinite period; and they are globally action-minimizing. When $\xi = 0$, the solution of the Mather problem is just the Dirac point on the unstable equilibrium x_0, and the Mather and Aubry sets Γ are reduced to $\{(x_0, x_0)\}$. When ξ varies in $\mathbb{R}$, this remains the same until ξ reaches a certain critical value; above that value, the Mather measures are supported by revolutions. At the critical value, the Mather and Aubry sets differ: the Aubry set (viewed in the variables (x, v)) is the union of the two revolutions of infinite period.

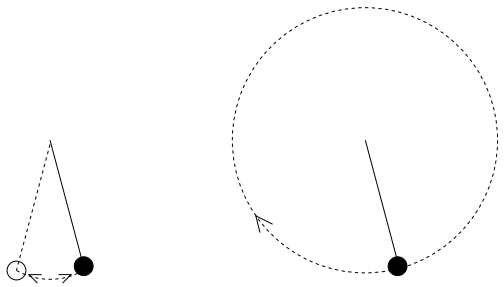


Fig. 8.5. On the left figure, the pendulum oscillates with little energy between two extreme positions; its trajectory is an arc of circle which is described clockwise, then counterclockwise, then clockwise again, etc. On the right figure, the pendulum has much more energy and draws complete circles again and again, either clockwise or counterclockwise.

The dual point of view in Mather's theory, and the notion of Aubry set, are intimately related to the so-called **weak KAM theory**, in which stationary solutions of Hamilton–Jacobi equations play a central role. The next theorem partly explains the link between the two theories.

Theorem 8.18 (Mather's theory and stationary solutions of Hamilton–Jacobi equations). *With the same notation as in Theorem 8.12, assume that the Lagrangian L does not depend on t, and let ψ be a Lipschitz function on M, such that $H_+^{0,t}\psi = \psi + ct$ for all times $t \geq 0$; that is, ψ is invariant by the forward Hamilton–Jacobi semigroup, except for the addition of a constant which varies linearly in time. Then, necessarily $c = \overline{c}$, and the pair $(\psi, H_+^{0,T}\psi) = (\psi, \psi + \overline{c}T)$ is optimal in the dual Kantorovich problem with measures $(\overline{\mu}, \overline{\mu})$ and cost function $c^{0,T}$.*

Remark 8.19. The equation $H_+^{0,1}\psi = \psi + ct$ is a way to reformulate the stationary Hamilton–Jacobi equation $H(x, \nabla\psi(x)) + c = 0$. Yet another reformulation would be obtained by changing the forward Hamilton–Jacobi for the backward one. Theorem 8.18 does not guarantee the existence of such stationary solutions, it just states that *if* such solutions exist, then the value of the constant c is uniquely determined and can be related to a Monge–Kantorovich problem. In weak KAM theory, one then establishes the existence of these solutions by independent means; see the references suggested in the bibliographical notes for much more information.

Proof of Theorem 8.18. To fix the ideas, let us impose $T = 1$. Let ψ be such that $H_+^{0,1}\psi = \psi + c$, and let μ be any probability measure on M; then

$$\int (H_+^{0,1}\psi)\, d\mu - \int \psi\, d\mu = \int c\, d\mu = c.$$

It follows from the easy part of the Kantorovich duality that $C^{0,1}(\mu,\mu) \geq c$. By taking the infimum over all $\mu \in P(M)$, we conclude that $\overline{c} \geq c$.

To prove the reverse inequality, it suffices to construct a particular probability measure μ such that $C^{0,1}(\mu,\mu) \leq c$. The idea is to look for μ as a limit of probability measures distributed uniformly over some well-chosen long minimizing trajectories. Before starting this construction, we first remark that since M is compact, there is a uniform bound C on $L(\gamma(t),\dot\gamma(t))$, for all action-minimizing curves $\gamma : [0,1] \to M$; and since L is time-independent, this statement trivially extends to all action-minimizing curves defined on time-intervals $[t_0,t_1]$ with $|t_0 - t_1| \geq 1$. Also ψ is uniformly bounded on M.

Let now x be an arbitrary point in M; for any $T > 0$ we have, by definition of the forward Hamilton–Jacobi semigroup,

$$(H_+^{-T,0}\psi)(x) = \inf\left\{\psi(\gamma(-T)) \ + \int_{-T}^0 L(\gamma(s),\dot\gamma(s))\, ds; \quad \gamma(0) = x\right\},$$

where the infimum is over all action-minimizing curves $\gamma : [-T,0] \to M$ ending at x. (The advantage to work with negative times is to fix one of the endpoints; in the present context where M is compact this is nonessential, but it would become important if M were noncompact.) By compactness, there is a minimizing curve $\gamma = \gamma^{(T)}$; then, by the definition of $\gamma^{(T)}$ and the stationarity of ψ,

$$\begin{aligned}\frac{1}{T}\int_{-T}^0 L\big(\gamma^{(T)}(s),\dot\gamma^{(T)}(s)\big)\, ds &= \frac{1}{T}\Big[(H_+^{-T,0}\psi)(x) - \psi(\gamma^{(T)}(-T))\Big] \\ &= \frac{1}{T}\Big(\psi(x) + cT - \psi(\gamma^{(T)}(-T))\Big) \\ &= c + O\left(\frac{1}{T}\right).\end{aligned}$$

In the sequel, I shall write just γ for $\gamma^{(T+1)}$. Of course the estimate above remains unchanged upon replacement of T by $T+1$, so

$$\frac{1}{T}\int_{-(T+1)}^0 L(\gamma^(s),\dot\gamma(s))\, ds = c + O\left(\frac{1}{T}\right).$$

Then define

$$\mu_T := \frac{1}{T}\int_{-(T+1)}^{-1} \delta_{\gamma(s)}\, ds; \qquad \nu_T := \frac{1}{T}\int_{-T}^0 \delta_{\gamma(s)}\, ds,$$

and $\theta : \gamma(s) \to \gamma(s+1)$. It is clear that $\theta_\# \mu_T = \nu_T$; moreover,

$$c^{0,1}(\gamma(s),\theta(\gamma(s))) = c^{0,1}(\gamma(s),\gamma(s+1)) = \int_s^{s+1} L(\gamma(u),\dot\gamma(u))\, du.$$

Thus

$$\begin{aligned}C^{0,1}(\mu_T,\nu_T) &\leq \frac{1}{T}\int_{-(T+1)}^{-1} c^{0,1}(\gamma(s),\theta(\gamma(s)))\, ds \\ &= \frac{1}{T}\int_{-(T+1)}^{-1}\left(\int_s^{s+1} L(\gamma(u),\dot\gamma(u))\, du\right) ds \\ &= \frac{1}{T}\int_{-(T+1)}^0 L(\gamma(u),\dot\gamma(u))\, a(u)\, du, \end{aligned} \tag{8.34}$$

where

$$a(u) = \int 1_{s \le u \le s+1}\, ds = \begin{cases} 1 & \text{if } -T \le u \le -1; \\ -u & \text{if } u \ge -1; \\ u + T + 1 & \text{if } u \le -T. \end{cases}$$

Replacing a by 1 in the integrand of (8.34) involves modifying the integral on a set of measure at most 2; so

$$C^{0,1}(\mu_T, \nu_T) \le \frac{1}{T} \int_{-(T+1)}^{0} L(\gamma(u), \dot\gamma(u))\, du + O\left(\frac{1}{T}\right) = c + O\left(\frac{1}{T}\right). \tag{8.35}$$

Since $P(M)$ is compact, the family $(\mu_T)_{T \in \mathbb{N}}$ converges, up to extraction of a subsequence, to some probability measure μ. Then (up to extraction of the same subsequence) ν_T also converges to μ, since

$$\|\mu_T - \nu_T\|_{TV} = \frac{1}{T} \Big\| \int_{-(T+1)}^{-T} \delta_{\gamma(s)}\, ds + \int_{-1}^{0} \delta_{\gamma(s)}\, ds \Big\|_{TV} \le \frac{2}{T}.$$

Then from (8.35) and the lower semi-continuity of the optimal transport cost,

$$C^{0,1}(\mu, \mu) \le \liminf_{T \to \infty} \; C^{0,1}(\mu_T, \nu_T) \; \le c.$$

This concludes the proof. □

The next exercise may be an occasion to manipulate the concepts introduced in this section.

Exercise 8.20. With the same assumptions as in Theorem 8.12, assume that L is symmetric in v; that is, $L(x, -v, t) = L(x, v, t)$. Show that $c^{0,T}(x, y) = c^{0,T}(y, x)$. Take an optimal measure $\overline{\mu}$ for the minimization problem (8.23), and let π be an associated optimal transference plan. By gluing together π and $\check\pi$ (obtained by exchanging the variables x and y), construct an optimal transference plan for the problem (8.23) with T replaced by $2T$, such that each point x stays in place. Deduce that the curves γ are $2T$-periodic. Show that $c^{0,2T}(x, x) = C^{0,2T}(\overline{\mu}, \overline{\mu})$, and deduce that $c^{0,T}(x, y)$ is π-almost surely constant. Construct ψ such that $H_+^{0,2T}\psi = \psi + 2\overline{c}T$, $\overline{\mu}$-almost surely. Next assume that L does not depend on t, and use a compactness argument to construct a ψ and a stationary measure μ, such that $H_+^{0,t}\psi = \psi + \overline{c}t$, for all $t \ge 0$, μ-almost surely. Note that Theorem 8.18 reinforces this result in two ways: It does not assume the symmetry of L (which is a huge simplification), and the equation $H_+^{0,t}\psi = \psi + \overline{c}t$ does not hold just μ-almost surely, but everywhere in M.

Possible extensions of Mather's estimates

As noticed in Example 8.4, it would be desirable to have a sharper version of Theorem 8.1 which would contain as a special case the correct exponents for the Lagrangian function $L(x, v, t) = |v|^{1+\alpha}$, $0 < \alpha < 1$.

But even for a "uniformly convex" Lagrangian there are several extensions of Theorem 8.1 which would be of interest, such as (a) getting rid of the compactness assumption; and (b) getting rid of the smoothness assumptions. I shall discuss both problems in the most typical case $L(x, v, t) = |v|^2$, i.e. $c(x, y) = d(x, y)^2$.

Intuitively, Mather's estimates are related to the behavior of geodesics (they should not diverge too fast), and to the convexity properties of the square distance function $d^2(x_0, \cdot)$. Both features are well captured by lower bounds on the **sectional curvature** of the manifold. There is by chance a generalized notion of sectional curvature bounds, due to Alexandrov, which makes sense in a general metric space, without any smoothness; metric spaces which satisfy these bounds are called **Alexandrov spaces**. (This notion will be explained in more detail in Chapter 26.) In such spaces, one could hope to solve problems (a) and (b) at the same time. Although the proofs in the present chapter strongly rely on smoothness, I would be ready to believe in the following statement (which might be not so difficult to prove):

Open Problem 8.21. *Let $(\mathcal{X}, d)$ be an Alexandrov space with curvature bounded below by $K \in \mathbb{R}$, and let x_1, x_2, y_1, y_2 be four points in $\mathcal{X}$ such that*

$$d(x_1, y_1)^2 + d(x_2, y_2)^2 \le d(x_1, y_2)^2 + d(x_2, y_1)^2.$$

Let further γ_1 and γ_2 be two constant-speed geodesics respectively joining x_1 to y_1 and x_2 to y_2. Then, for any $t_0 \in (0,1)$, there is a constant C_{t_0}, depending only on K and t_0, and maybe on an upper bound on all the distances involved, such that

$$\sup_{0 \le t \le 1} d\big(\gamma_1(t), \gamma_2(t)\big) \le C_{t_0}\, d\big(\gamma_1(t_0),\, \gamma_2(t_0)\big).$$

To conclude this discussion, I shall mention a much rougher "shortening lemma", which has the advantage to hold true in general metric spaces, even without curvature bounds. In such a situation, in general there may be branching geodesics, so a bound on the distance at one intermediate time is clearly not enough to control the distance between the positions along the whole geodesic curves. One cannot hope either to control the distance between the velocities of these curves, since the velocities might not be well-defined. On the other hand, we may take advantage of the property of *preservation of speed* along the minimizing curves, since this remains true even in a nonsmooth context. The next theorem exploits this to show that if geodesics in a displacement interpolation pass nearby each other at some intermediate time, then their *lengths* have to be approximately equal.

Theorem 8.22. *Let $(\mathcal{X}, d)$ be a metric space, and let γ_1, γ_2 be two constant-speed, minimizing geodesics such that*

$$d\big(\gamma_1(0), \gamma_1(1)\big)^2 + d\big(\gamma_2(0), \gamma_2(1)\big)^2 \le d\big(\gamma_1(0), \gamma_2(1)\big)^2 + d\big(\gamma_2(0), \gamma_1(1)\big)^2.$$

Let L_1 and L_2 stand for the respective lengths of γ_1 and γ_2, and let D be a bound on the diameter of $(\gamma_1 \cup \gamma_2)([0,1])$. Then

$$|L_1 - L_2| \le \frac{C\sqrt{D}}{\sqrt{t_0(1-t_0)}}\, \sqrt{d\big(\gamma_1(t_0), \gamma_2(t_0)\big)},$$

for some numeric constant C.

Proof. Write $d_{12} = d(x_1, y_2)$, $d_{21} = d(x_2, y_1)$, $X_1 = \gamma_1(t_0)$, $X_2 = \gamma_2(t_0)$. From the minimizing assumption, the triangular inequality and explicit calculations,

$$\begin{aligned}
0 &\le d_{12}^2 + d_{21}^2 - L_1^2 - L_2^2 \\
&\le \Big(t_0\, L_1 + d(X_1, X_2) + (1-t_0)\, L_2\Big)^2 + \Big(t_0\, L_2 + d(X_1, X_2) + (1-t_0)\, L_1\Big)^2 - L_1^2 - L_2^2 \\
&= 2\, d(X_1, X_2)\Big(L_1 + L_2 + d(X_1, X_2)\Big) - 2\, t_0(1-t_0)\, (L_1 - L_2)^2.
\end{aligned}$$

As a consequence,

$$|L_1 - L_2| \le \sqrt{\frac{L_1 + L_2 + d(X_1, X_2)}{t_0(1-t_0)}}\ \sqrt{d(X_1, X_2)},$$

as desired. □

Appendix: Lipschitz estimates for power cost functions

The goal of this Appendix is to prove the following shortening lemma for the cost function $c(x,y) = |x-y|^{1+\alpha}$ in Euclidean space.

Theorem 8.23 (Shortening lemma for power cost functions). *Let $\alpha \in (0,1)$, and let x_1, y_1, x_2, y_2 be four points in $\mathbb{R}^n$, such that*

$$|x_1 - y_1|^{1+\alpha} + |x_2 - y_2|^{1+\alpha} \le |x_1 - y_2|^{1+\alpha} + |x_2 - y_1|^{1+\alpha}. \tag{8.36}$$

Let further

$$\gamma_1(t) = (1-t)\, x_1 + t\, y_1, \qquad \gamma_2(t) = (1-t)\, x_2 + t\, y_2.$$

Then, for all $t_0 \in (0,1)$ there is a constant $K = K(\alpha, t_0) > 0$ such that

$$|\gamma_1(t_0) - \gamma_2(t_0)| \ge K \sup_{0 \le t \le 1} |\gamma_1(t) - \gamma_2(t)|.$$

Remark 8.24. The proof below is not constructive, so I won't have any quantitative information on the best constant $K(\alpha, t)$. It is natural to think that for each fixed t, the constant $K(\alpha, t)$ (which only depends on α) will go to 0 as $\alpha \downarrow 0$. When $\alpha = 0$, the conclusion of the Proposition is false: Just think of the case when x_1, y_1, x_2, y_2 are aligned. But this is the only case where the conclusion fails, so it might be that a modified statement still holds true.

Proof of Theorem 8.23. First note that it suffices to work in the affine space generated by x_1, y_1, x_2, y_2, which is of dimension at most 3; hence all the constants will be independent of the dimension n. For notational simplicity, I shall assume that $t_0 = 1/2$, which has no important influence on the computations. Let $X_1 := \gamma_1(1/2)$, $X_2 := \gamma_2(1/2)$. It is sufficient to show that

$$|x_1 - x_2| + |y_1 - y_2| \le C\, |X_1 - X_2|$$

for some constant C, independent of x_1, x_2, y_1, y_2.

Step 1: Reduction to a compact problem by invariance. Exchanging the roles of x and y, we might assume that $|x_2 - y_2| \le |x_1 - y_1|$, and then by translation invariance that $x_1 = 0$, by homogeneity that $|x_1 - y_1| = 1$ (treat separately the trivial case $x_1 = y_1$), and by rotation invariance that $y_1 = e$ is a fixed unit vector.

Let $R := |x_2|$, then $|y_2 - x_2| \le 1$ implies that $|X_2| \ge R - 1/2$, and since $|X_1| \le 1/2$, it follows that $|X_1 - X_2| \ge R - 1$. Similarly, $|x_1 - x_2| \le R + 1$ and $|y_1 - y_2| \le R + 2$. So the conclusion is obvious if $R \ge 2$. Otherwise, $|x_2|$ and $|y_2|$ lie in the ball $B_3(0)$.

Step 2: Reduction to a perturbation problem by compactness. For any positive integer k, let $(x_2^{(k)}, y_2^{(k)})$ be such that $(|x_1 - x_2| + |y_1 - y_2|)/|X_1 - X_2|$ is minimized by $(x_1, y_1, x_2^{(k)}, y_2^{(k)})$ under the constraint $|X_1 - X_2| \ge k^{-1}$.

By compactness, such a configuration does exist, and the value I_k of the infimum goes down with k, and converges to

$$I := \inf \left(\frac{|x_1 - x_2| + |y_1 - y_2|}{|X_1 - X_2|} \right), \tag{8.37}$$

where the infimum is taken over all configurations such that $X_1 \neq X_2$. However, from the strict convexity of $|x|^{1+\alpha}$ it follows that inequality (8.36) prevents $X_1 = X_2$, unless $(x_1, y_1) = (x_2, y_2)$, in which case there is nothing to prove. So it is really sufficient to show that $I > 0$.

Since the sequence $(x_2^{(k)}, y_2^{(k)})$ takes values in a compact set, there is a subsequence thereof (still denoted $(x_2^{(k)}, y_2^{(k)})$) which converges to some $(x_2^{(\infty)}, y_2^{(\infty)})$. By continuity, condition (8.36) holds true with $(x_2, y_2) = (x_2^{(\infty)}, y_2^{(\infty)})$. If one has (with obvious notation) $|X_1 - X_2^{(\infty)}| > 0$, then the configuration $(x_1, y_1, x_2^{(\infty)}, y_2^{(\infty)})$ achieves the minimum I in (8.37), and that minimum is positive. So the only case that remains to treat is when $X_2^{(\infty)} = X_1$. Then, by strict convexity, condition (8.36) imposes $x_2^{(\infty)} = x_1$, $y_2^{(\infty)} = y_1$. Equivalently, $x_2^{(k)}$ converges to x_1, and $y_2^{(k)}$ to y_1. All this shows that it suffices to treat the case when x_2 is very close to x_1 and y_2 is very close to y_1.

Step 3: Expansions. Now let

$$x_2 = x_1 + \delta x, \qquad y_2 = y_1 + \delta y, \tag{8.38}$$

where δx and δy are vectors of small norm (recall that $x_1 - y_1$ has unit norm). Of course

$$X_1 - X_2 = \frac{\delta x + \delta y}{2}, \qquad x_1 - x_2 = \delta x, \qquad y_1 - y_2 = \delta y;$$

so to conclude the proof it is sufficient to show that

$$\left| \frac{\delta x + \delta y}{2} \right| \geq K(|\delta x| + |\delta y|), \tag{8.39}$$

as soon as $|\delta x|$ and $|\delta y|$ are small enough, and (8.36) is satisfied.

By using the formulas $|a + b|^2 = |a|^2 + 2\langle a, b\rangle + |b|^2$ and

$$(1 + \varepsilon)^{\frac{1+\alpha}{2}} = 1 + \frac{(1+\alpha)}{2}\,\varepsilon \; - \frac{(1+\alpha)(1-\alpha)}{8}\varepsilon^2 + O(\varepsilon^3),$$

one easily deduces from (8.36) that

$$|\delta x - \delta y|^2 - |\delta x|^2 - |\delta y|^2 \leq (1-\alpha)\Big[\langle \delta x - \delta y, e\rangle^2 - \langle \delta x, e\rangle^2 - \langle \delta y, e\rangle^2\Big] + O\Big(|\delta x|^3 + |\delta y|^3\Big).$$

This can be rewritten

$$\langle \delta x, \delta y\rangle - (1-\alpha)\langle \delta x, e\rangle\,\langle \delta y, e\rangle \geq O(|\delta x|^3 + |\delta y|^3).$$

Consider the new scalar product

$$\langle\!\langle v, w\rangle\!\rangle := \langle v, w\rangle - (1-\alpha)\langle v, e\rangle\,\langle w, e\rangle$$

(which is indeed a scalar product because $\alpha > 0$), and denote the associated norm by $\|v\|$. Then the above conclusion can be summarized into

$$\langle\!\langle \delta x, \delta y\rangle\!\rangle \geq O\big(\|\delta x\|^3 + \|\delta y\|^3\big). \tag{8.40}$$

It follows that

$$\left\| \frac{\delta x + \delta y}{2} \right\|^2 = \frac{1}{4}\Big(\|\delta x\|^2 + \|\delta y\|^2 + 2\langle\!\langle \delta x, \delta y \rangle\!\rangle \Big)$$
$$\geq \frac{1}{4}(\|\delta x\|^2 + \|\delta y\|^2) + O(\|\delta x\|^3 + \|\delta y\|^3).$$

So inequality (8.39) is indeed satisfied if $|\delta x| + |\delta y|$ is small enough. □

Exercise 8.25. Extend this result to the cost function $d(x,y)^{1+\alpha}$ on a Riemannian manifold, when γ and $\widetilde{\gamma}$ stay within a compact set. *Hints:* This is a difficult exercise, only for a reader that feels very comfortable. One can use a reasoning similar to that in Step 2 of the above proof, introducing a sequence $(\gamma^{(k)}, \widetilde{\gamma}^{(k)})$ which is asymptotically "worst possible", and converges, up to extraction of a subsequence, to $(\gamma^{(\infty)}, \widetilde{\gamma}^{(\infty)})$. There are three cases: (i) $\gamma^{(\infty)}$ and $\widetilde{\gamma}^{(\infty)}$ are distinct geodesic curves which cross; this is ruled out by Theorem 8.1. (ii) $\gamma^{(k)}$ and $\widetilde{\gamma}^{(k)}$ converge to a point; then everything becomes local and one can use the result in $\mathbb{R}^n$, Theorem 8.23. (iii) $\gamma^{(k)}$ and $\widetilde{\gamma}^{(k)}$ converge to a nontrivial geodesic $\gamma^{(\infty)}$; then these curves can be approximated by infinitesimal perturbations of $\gamma^{(\infty)}$, which are described by differential equations (Jacobi equation).

Remark 8.26. Of course it would be much better to avoid the compactness arguments and derive the bounds directly, but I don't see how to proceed.

Bibliographical Notes

Monge's observation about the impossibility of crossing appears in his seminal 1781 memoir [276]. The argument is likely to apply whenever the cost function satisfies a triangle inequality, which is always the case in what Bernard and Buffoni have called the Monge–Mañé problem [48]. I don't know of a quantitative version of it.

A very simple argument, due to Brenier, shows how to construct, without any calculations, configurations of points that lead to line-crossing for a quadratic cost [365, Chapter 10, Problem 1].

There are several possible computations to obtain inequalities of the style of (8.3). The use of the identity (8.2) is inspired from a result by Figalli, which is described below.

Mather's shortening lemma was published in 1991 [260, p. 186]; it was the key technical estimate in the proof of his "Lipschitz Graph Theorem" [260, Theorem 2]. Theorem 8.12 is a variant of Mather's theorem, appearing (up to minor modifications) in a recent work by Bernard and Buffoni [47, Theorem C]. The core of the proof is also taken from that work.

The "weak KAM theory" was developed by several authors, in particular Fathi [**?**, **?**]. The reader can also consult the lecture notes [160], or the review works [330, 169].

The proof of Theorem 8.18, as I wrote it, is a minor variation of an argument shown to me by Fathi.

From its very beginning, the weak KAM theory has been associated with the theory of viscosity solutions of Hamilton–Jacobi equations. An early work on the subject (anterior to Mather's papers) is an unpublished preprint by P.-L. Lions, Papanicolaou and Varadhan [243]. Recently, the weak KAM theory has been related to the large-time behavior of Hamilton–Jacobi equations [**?**, **?**, **?**, **?**]. Aubry sets are also related with the C^1 regularity of Hamilton–Jacobi equations [**?**, **?**, **?**]. See also Evans and Gomes [157, 158, **?**, **?**] and the references therein for an alternative point of view.

In this chapter I presented Mather's problem in terms of trajectories and transport cost. There is an alternative presentation in terms of invariant measures, following an idea by

Mañé. In Mañé's version of the problem, the unknown is a probability measure $\mu(dx\,dv)$ on the tangent bundle TM; it is stationary in the sense that $\nabla_x(v\,\mu) = 0$ (this is a stationary kinetic transport equation), and it should minimize the action $\int L(x,v)\,\mu(dx\,dv)$. Then one can show that μ is actually invariant under the Lagrangian flow defined by L. As Gomes pointed out to me, this approach has the drawback that the invariance of μ is not built in from the definition; but it has several nice advantages:

- it makes the graph theorem trivial if L is strictly convex: Indeed, one can always collapse the measure μ, at each $x \in M$, onto the barycenter $\xi(x) = \int v\,\mu(dv|x)$; this operation preserves the invariance of the measure, and decreases the cost unless μ was already supported on a graph.

- this is a linear programming problem, with dual problem $\inf_\varphi \sup_x H(\nabla_x\varphi, x)$; the value of this infimum is but another way to characterize the effective Hamiltonian $\overline{H}$ (see [**?**]).

- this is a good starting point for some generalizations, see for instance [**?**].

The "no-crossing" property of optimal trajectories, and the resulting estimates about absolute continuity of the displacement interpolant, were some of the key technical tools used by McCann [267] to establish convexity properties of certain functionals along displacement interpolation in $\mathbb{R}^n$ for a quadratic cost. Later this was generalized to Riemannian manifolds by Cordero-Erausquin, McCann and Schmuckenschläger [118]; Cordero-Erausquin [116] also adapted the techniques of the latter paper to treat rather general convex cost functions in Euclidean space. More recently, Bernard and Buffoni suggested that the use of Mather's lemma could simplify and generalize these estimates; this is the approach which I have implemented in these notes. Bernard and Buffoni themselves preferred to base their proofs on the theory of Hamilton–Jacobi equations, which is less elementary.

The use of a restriction property to prove the absolute continuity of the displacement interpolant without any compactness assumption was inspired by a discussion with Sturm on a related subject. It was also Sturm who asked me whether Mather's estimates could be generalized to Alexandrov spaces with curvature bounded below.

The theorem according to which a Lipschitz map T dilates the n-dimensional Hausdorff measure by a factor at most $\|T\|^n_{\mathrm{Lip}}$ is an almost immediate consequence of the definitions of Hausdorff measure, see e.g. [81, Proposition 1.7.8].

Alexandrov spaces are discussed at length in the very pedagogical monograph by Burago, Burago and Ivanov [81]. Several characterizations of Alexandrov spaces are given there, and their equivalence is established. For instance, an Alexandrov space has curvature bounded below by K if the square distance function $d(z,\cdot)^2$ is "no more convex" than the square distance function in the model space having constant sectional curvature K. Also geodesics in an Alexandrov space cannot diverge faster than geodesics in the model space, in some sense. These properties explain why it is possible to believe that such spaces are a natural generalized setting for optimal transport. Upper bounds on the sectional curvature, on the other hand, do not seem to be of any help.

Figalli recently solved the Open Problem 8.21 in the special case $K = 0$ (nonnegative curvature), with a very simple and sharp argument. He actually showed that if γ_1 and γ_2 are any two minimizing, constant-speed geodesics in an Alexandrov space $(\mathcal{X}, d)$ with nonnegative curvature, and $\gamma_1(0) = x_1$, $\gamma_2(0) = x_2$, $\gamma_1(1) = y_1$, $\gamma_2(1) = y_2$, then

$$d(\gamma_1(t),\,\gamma_2(t)) \geq (1-t)^2\,d(x_1,x_2)^2 \,+\, t^2\,d(y_1,y_2)^2 \\ +\, t(1-t)\left[d(x_1,y_2)^2 + d(x_2,y_1)^2 - d(x_1,y_1)^2 - d(x_2,y_2)^2\right]. \quad (8.41)$$

The general case where K might be negative seems to be quite more tricky.

Theorem 8.22 takes inspiration from the no-crossing argument in [118]. I don't know whether the Hölder-1/2 regularity is optimal, and I don't know either whether it is possible/useful to obtain similar estimates for more general cost functions.

9

Solution of the Monge problem, I (Global approach)

In the present chapter and the next one I shall investigate the solvability of the Monge problem for a Lagrangian cost function. Recall from Theorem 5.25 that it is sufficient to identify conditions under which the initial measure μ does not see the set of points where the subdifferential of a c-convex function ψ is multivalued.

Consider a Riemannnian manifold M, and a cost function $c(x, y)$ on $M \times M$, deriving from a Lagrangian function $L(x, v, t)$ on $TM \times [0, 1]$. Let two probability measures μ_0 and μ_1 be given, and let $(\mu_t)_{0 \le t \le 1}$ be a displacement interpolation, written as the law of a random minimizing curve γ at time t.

If the Lagrangian satisfies adequate regularity and convexity properties, Theorem 8.5 shows that the coupling $(\gamma(s), \gamma(t))$ is always deterministic, as soon as $0 < s < 1$, however singular μ_0 and μ_1 might be. The question whether one can construct a deterministic coupling of (μ_0, μ_1) is much more subtle, and cannot be answered without regularity assumptions on μ_0. In this chapter, a simple approach of this problem will be attempted, but only with partial success, since eventually it will work out only for a particular class of cost functions, including at least the quadratic cost in Euclidean space (arguably the most important case).

Our main assumption on the cost function c will be the following.

Assumption (C): *For any c-convex function ψ and any $x \in M$, the c-subdifferential $\partial_c \psi(x)$ is pathwise connected.*

Example 9.1. Consider the cost function $c(x, y) = -x \cdot y$ in $\mathbb{R}^n$. Let y_0 and y_1 belong to $\partial_c \psi(x)$; then, for all $z \in \mathbb{R}^n$ one has

$$\psi(x) + y_0 \cdot (z - x) \le \psi(z); \qquad \psi(x) + y_1 \cdot (z - x) \le \psi(z).$$

It follows that $\psi(x) + y_t \cdot (z - x) \le \psi(z)$, where $y_t := (1 - t)y_0 + t y_1$. Thus the line segment $(y_t)_{0 \le t \le 1}$ is entirely contained in the subdifferential of ψ at x. The same computation applies to $c(x, y) = |x - y|^2/2$, or to any cost function of the form $a(x) - x \cdot y + b(y)$.

Actually, there are few examples where Assumption (C) is known to be satisfied. Before commenting more on that issue, let me illustrate the interest of this assumption by showing how it can be used.

Theorem 9.2 (Conditions for single-valued subdifferentials). *Let M be a smooth n-dimensional manifold, and c a real-valued cost function, with c bounded below, deriving from a Lagrangian cost function $L(x, v, t)$ on $TM \times [0, 1]$, such that*

(i) Assumption (C) is satisfied.

(ii) The conclusion of Theorem 8.1 (Mather's shortening lemma), in the form of inequality (8.4)*, holds true for* $t_0 = 1/2$ *with an exponent* $\beta > 1 - (1/n)$*, and a uniform constant. More explicitly: Whenever* x_1, x_2, y_1, y_2 *are four points on* M *satisfying* $c(x_1, y_1) + c(x_2, y_2) \le c(x_1, y_2) + c(x_2, y_1)$*, and* γ_1*,* γ_2 *are two action-minimizing curves with* $\gamma_1(0) = x_1$*,* $\gamma_1(1) = y_1$*,* $\gamma_2(0) = x_2$*,* $\gamma_2(1) = y_2$*, then*

$$\sup_{0 \le t \le 1} d\big(\gamma_1(t), \gamma_2(t)\big) \le C\, d\big(\gamma_1(1/2),\, \gamma_2(1/2)\big). \tag{9.1}$$

Then, for any c-convex function ψ*, there is a set* $Z \subset M$ *of Hausdorff dimension at most* $(n-1)/\beta < n$ *(and therefore of zero n-dimensional measure), such that the c-subdifferential* $\partial_c \psi(x)$ *contains at most one element if* $x \notin Z$*.*

Proof. Let Z be the set of points x for which $\psi(x) < +\infty$ but $\partial_c \psi(x)$ is not single-valued; the problem is to show that Z is of dimension at most $(n-1)/\beta$.

Let $x \in M$ with $\psi(x) < +\infty$, and let $y \in \partial_c \psi(x)$. Introduce an action-minimizing curve $\gamma = \gamma^{x,y}$ joining $x = \gamma(0)$ to $y = \gamma(1)$. I claim that the map

$$F : \ \gamma\left(\frac{1}{2}\right) \longmapsto x$$

is well-defined on its domain of definition, which is the union of all $\gamma^{x,y}(1/2)$. (I mean, $m = \gamma(1/2)$ determines x unambiguously; there cannot be two different points x for which $\gamma(1/2)$ is the same.) Indeed, assume $y \in \partial_c \psi(x)$ and $y' \in \partial_c \psi(x')$, with $\psi(x) < +\infty$, $\psi(x') < +\infty$, and let γ and γ' be minimizing geodesics between x and y on one hand, x' and y' on the other hand. It follows from the definitions of subdifferential that

$$\begin{cases} \psi(x) \ + c(x, y) \ \le \psi(x') \ + c(x', y) \\ \psi(x') + c(x', y') \le \psi(x) + c(x, y) \end{cases}$$

Thus

$$c(x, y) + c(x', y') \le c(x, y') + c(x', y).$$

Then Theorem 8.1 implies that

$$d(x, x') \le C\ d\Big(\gamma(\frac{1}{2}),\, \gamma'(\frac{1}{2})\Big)^{\beta}.$$

If follows that $m = \gamma(1/2)$ determines $x = F(m)$ unambiguously, and even that F is Hölder-β. (Obviously, this is the same reasoning as in Theorem 8.5.)

Now, cover M by a countable number of open sets in which M is diffeomorphic to a subset U of $\mathbb{R}^n$, via some diffeomorphism φ_U. In each of these open sets U, consider the union H_U of all hyperplanes passing through a rational point, orthogonal to a unit vector with rational coordinates. Transport this set back to M thanks to the local diffeomorphism; now take the union over all the sets U. This gives a set $D \subset M$ with the following properties: (i) It is of dimension $n-1$; (ii) It meets every nontrivial continuous curve drawn on M (to see this, write the curve locally in terms of φ_U and note that, by continuity, at least one of the coordinates of the curve has to become rational at some time).

Next, let $x \in Z$, and let y_0, y_1 be two distinct elements of $\partial_c \psi(x)$. By assumption there is a continuous curve $(y_t)_{0 \le t \le 1}$ lying entirely in $\partial_c \psi(x)$. For each t, introduce an action-minimizing curve $(\gamma_t(s))_{0 \le s \le 1}$ between x and y_t (s here is the time parameter along the curve). Define $m_t := \gamma_t(1/2)$. This is a continuous path, nontrivial (otherwise

$\gamma_0(1/2) = \gamma_1(1/2)$, but two minimizing trajectories starting from x cannot cross in their middle, otherwise they have to coincide for all times by (9.1)). So there has to be some t such that $y_t \in D$. Moreover, the map F constructed above is constant on D: $F(y_t) = x$ for all t. It follows that $x \in F(D)$.

As a conclusion, $Z \subset F(D)$. Since D is of Hausdorff dimension $n-1$ and F is β-Hölder, it follows that the dimension of $F(D)$ is at most $(n-1)/\beta$. □

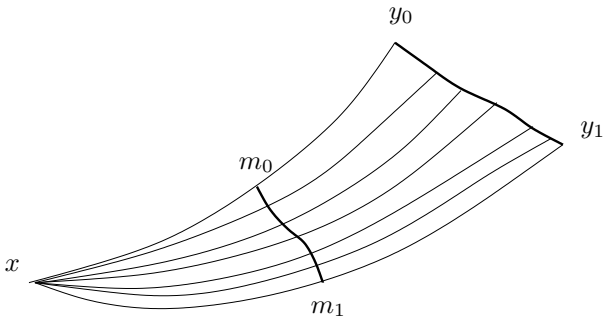


Fig. 9.1. Scheme of proof for Proposition 9.2. Here there is a curve $(y_t)_{0\le t\le 1}$ lying entirely in $\partial_c\psi(x)$, and there is a nontrivial path $(m_t)_{0\le t\le 1}$ obtained by taking the midpoint between x and y_t. This path has to meet D; but its image by $\gamma(1/2) \to \gamma(0)$ is $\{x\}$, so $x \in F(D)$.

Now come the consequences in terms of Monge transport.

Corollary 9.3 (Solution of the Monge problem, I). *Let M be a Riemannian manifold, let c be a cost function on $M \times M$, with associated cost functional C, and let μ, ν be two probability measures on M. Assume that*

(i) $C(\mu,\nu) < +\infty$;

(ii) The assumptions of Theorem 9.2 are satisfied;

(iii) μ attributes zero probability to sets of dimension at most $(n-1)/\beta$.

Then, there exists a unique (in law) optimal coupling (x,y) of μ and ν; it is deterministic, and characterized (among all couplings of (μ,ν)) by the existence of a c-convex function ψ such that

$$y \in \partial_c\psi(x) \qquad \text{almost surely.} \tag{9.2}$$

Equivalently, there is a unique optimal transport plan π; it is deterministic, and characterized by the existence of a c-convex ψ such that (9.2) *holds true π-almost surely.*

Proof of Corollary 9.3. The conclusion is obtained by just putting together Theorems 9.2 and 5.25. □

Now we have solved the Monge problem in an absolutely painless way; but under what assumptions? It is a frustrating open problem that Assumption (C), simple as it may seem, is not known to be true for rather general cost functions. In fact, the realization of this condition seems to involve subtle features of the cost function, and is probably false in general; see the bibliographical notes for more details. The only case in which we can conclude, right now, is the cost function $c(x,y) = -x \cdot y$. For that cost function the notion of c-convexity reduces to plain convexity (plus lower semi-continuity), and the c-subdifferential of a convex function ψ is just its usual subdifferential, so it will be denoted by $\partial\psi$. Moreover, under an assumption of finite second moments, for the Monge problem

this cost is just as good as the quadratic Euclidean distance, since $|x-y|^2 = |x|^2 - 2x\cdot y + |y|^2$, and $\int (|x|^2 + |y|^2)\, d\pi(x,y)$ is independent of the choice of $\pi \in \Pi(\mu,\nu)$. So at present we are able to solve the case of the square Euclidean distance under an assumption of finite second moments. Since this is still one of the most important cases for applications, I shall state the result as a separate theorem.

Theorem 9.4 (Monge problem for quadratic cost, first result). *Let $c(x,y) = |x-y|^2$ in $\mathbb{R}^n$. Let μ, ν be two probability measures on $\mathbb{R}^n$ such that*

$$\int |x|^2\, d\mu(x) + \int |y|^2\, d\nu(y) < +\infty \tag{9.3}$$

and μ does not give mass to sets of dimension at most $n-1$. (This is true in particular if μ is absolutely continuous with respect to the Lebesgue measure.) Then there is a unique (in law) optimal coupling (x,y) of μ and ν; it is deterministic, and characterized (among all couplings of (μ,ν)) by the existence of a lower semi-continuous convex function ψ such that

$$y \in \partial\psi(x) \qquad \text{almost surely} \tag{9.4}$$

for some lower semi-continuous convex function ψ. In other words, there is a unique optimal transference π; it is a Monge transport plan, and it is characterized by the existence of a lower semi-continuous convex function ψ whose subdifferential contains $\mathrm{Spt}\,\pi$.

Remark 9.5. The assumption that μ does not give mass to sets of dimension at most $n-1$ is optimal for the existence of a Monge coupling, as can be seen by choosing $\mu = \mathcal{H}^1|_{\{0\}\times[0,1]}$ (the one-dimensional Hausdorff measure concentrated on the segment $\{0\}\times[0,1]$ in $\mathbb{R}^2$), and $\nu = (1/2)\mathcal{H}^1|_{\{-1\}\times[0,1]\cup\{+1\}\times[0,1]}$. It is also optimal for the uniqueness, as can be seen by taking $\mu = (1/2)\mathcal{H}^1_{\{0\}\times[-1,1]}$ and $\nu = (1/2)\mathcal{H}^1_{[-1,1]\times\{0\}}$. In fact, whenever μ, $\nu \in P_2(\mathbb{R}^n)$ are supported on orthogonal subspaces of $\mathbb{R}^n$, then *any* transference plan is optimal! To see this, define a convex function ψ by $\psi = 0$ on $\mathrm{Conv}(\mathrm{Spt}\,\mu)$, $\psi = +\infty$ elsewhere; then $\psi^* = 0$ on $\mathrm{Conv}(\mathrm{Spt}\,\nu)$, so $\partial\psi$ contains $\mathrm{Spt}\,\mu \times \mathrm{Spt}\,\nu$, and any transference plan is contained in $\partial\psi$.

Fig. 9.2. The source measure is drawn in thick line, the target measure in thin line; the cost function is quadratic. On the left, there is a unique optimal coupling but no optimal Monge coupling. On the right, there are many optimal couplings, in fact any transference plan is optimal.

In the next chapter, we shall see that Theorem 9.4 can be improved in at least two ways: The equation (9.4) can be rewritten $y = \nabla\psi(x)$; and the assumption (9.3) can be replaced by the weaker assumption $C(\mu,\nu) < +\infty$ (finite optimal transport cost).

Bibliographical Notes

It is classical that the image of a set of Hausdorff dimension d by a Lipschitz map is contained in a set of Hausdorff dimension at most d: See for instance [156, p. 75]. There is no difficulty in modifying the proof to show that the image of a set of Hausdorff dimension d by a Hölder-β map is contained in a set of dimension at most d/β.

The proof of Theorem 9.2 is adapted from a classical argument according to which a real-valued convex function ψ on $\mathbb{R}^n$ has a single-valued subdifferential everywhere out of a set of dimension at most $n-1$; see [4]. The key estimate for the proof of the latter theorem is that $(\mathrm{Id} + \partial\psi)^{-1}$ exists and is Lipschitz; but this can be seen as a very particular case of the Mather shortening lemma. In the next chapter another line of argumentation for that differentiability theorem, more local, will be provided.

Assumption (C) is alluded to briefly in a paper by Ma, Trudinger and Wang [250, Section 7.5]. For cost functions of the form $c(x-y)$, where c is convex on $\mathbb{R}^n$, these authors suggested that it is related to a certain complicated condition involving fourth-order derivatives of c, which is useful (and possibly mandatory) for the regularity of c-convex functions arising in optimal transport. They further conjectured that $c(x,y) = |x-y|^p$ should satisfy Assumption (C) for $p \in [1,2]$; but they proved that it does not for $p > 2$. If it is true that Assumption (C) is satisfied for $p \in (1,2)$, then this can be combined with Proposition 8.23 to immediately extend Theorem 9.4 to such cost functions, with obvious changes:

- choose $c(x,y) = |x-y|^p/p$ (say), $1 < p < 2$;
- replace moments of order 2 by moments of order p;
- replace the equation $y \in \partial\psi(x)$ by $y \in \partial_c\psi(x)$, where now ψ is c-convex.

Theorems of unique solvability of the Monge problem for such cost functions were proven long ago by Gangbo and McCann [179], with a different method.

A condition which is stronger than Assumption (C) is that $\partial_c\psi(x)$ should be c-convex [246]. This roughly means the following: The map $y \to -\nabla_x c(x,y)$ should be injective, and if F_x stands for its inverse, then for all $y, y' \in \partial_c\psi(x)$,

$$\forall t \in [0,1] \qquad F_x\big(-(1-t)\nabla_x c(x,y) - t\nabla_x c(x,y')\big) \in \partial_c\psi(x).$$

It was recently proved by Loeper [246] that this condition is essentially equivalent to the condition suggested by Ma, Trudinger and Wang, and also that it is essentially mandatory to develop a regularity theory for optimal transport. But it seems very unlikely that the condition of c-convexity of the c-subdifferentials is generic.

Loeper also managed to prove that Assumption (C) is satisfied when $c(x,y)$ is the squared geodesic distance on the Riemannian sphere. Combining this with Mather's estimates (Theorem 8.1), we can easily adapt the proof of Theorem 9.4 into a theorem of unique solvability of the Monge problem for the quadratic distance on the sphere, as soon as μ does not see sets of dimension at most $n-1$. Such a theorem was first obtained by McCann [269], with a completely different argument.

It might still be, that the proof of Theorem 9.3 can be cleverly modified to treat cases where Assumption (C) is not necessarily satisfied. But so far the scheme of proof only applies to very specific cases, in contrast with the method that will be presented in the next chapter.

The paternity of Theorem 9.4 is shared by Brenier [69, 72] on one hand, Rachev and Rüschendorf [322] on the other hand; it builds upon earlier work by Knott and Smith [229], who already knew that an optimal coupling lying entirely in the subdifferential of a convex function would be optimal. Brenier rewrote the result as a beautiful **polar factorization theorem**, which is presented in detail in [365, Chapter 3].

The nonuniqueness statement in Remark 9.5 was formulated by McCann [266]. Related problems (existence and uniqueness of optimal couplings between measures supported on polygons) are discussed by Gangbo and McCann [**?**], in relation with problems of shape recognition.

Other forms of Theorem 9.4 appear in Rachev and Rüschendorf [306], in particular an extension to an infinite-dimensional case (Hilbert spaces); the proof is reproduced in [365, Second Proof of Theorem 2.9]. All these proofs are based on duality; then more direct proofs, which do not use the Kantorovich duality explicitly, were found by Gangbo [176], and also Caffarelli [90] (who gives credit to Varadhan for this approach).

A probabilistic approach of Theorem 9.4 was studied by Mikami and Thieullen [**?**]. The idea is to consider a minimization problem over paths which are not geodesics, but geodesics perturbed by some noise; then to let the noise vanish. This is related to Nelson's approach of quantum mechanics, which I briefly mentioned in the bibliographical notes of Chapter 7.

McCann [266] extended Theorem 9.4 by removing the assumption of bounded second moment and even the weaker assumption of finite transport cost: *Whenever μ does not charge sets of dimension $n-1$, there exists a unique coupling of (μ,ν) which takes the form $y = \nabla\Psi(x)$, where Ψ is a lower semi-continuous convex function.* The tricky part in this statement is the uniqueness. This theorem will be proven in the next chapter (see Theorem 10.36 and Remark **??**).

10

Solution of the Monge problem, II (Local approach)

In the previous chapter, we tried to establish the almost sure single-valuedness of the c-subdifferential by an argument involving "global" topological properties, such as connectedness. Since this strategy worked out only in certain particular cases, we shall now explore a different approach, based on *local* properties of c-convex functions. The idea is that the global question *"Is the c-subdifferential of ψ at x single-valued or not?"* might be much more subtle to attack than the local question *"Is the function ψ differentiable at x or not?"* For a large class of cost functions, these questions are in fact equivalent; but these different formulations suggest different strategies. So in this chapter, the emphasis will be on tangent vectors and gradients, rather than elements of the c-subdifferential.

A heuristic argument

Let ψ be a c-convex function on a Riemannian manifold M, and $\phi = \psi^c$. Assume that $y \in \partial_c \psi(x)$; then, from the definition of c-subdifferential, one has, for all $\widetilde{x} \in M$,

$$\begin{cases} \phi(y) - \psi(x) = c(x,y) \\ \phi(y) - \psi(\widetilde{x}) \leq c(\widetilde{x},y). \end{cases} \tag{10.1}$$

It follows that

$$\psi(x) - \psi(\widetilde{x}) \leq c(\widetilde{x},y) - c(x,y). \tag{10.2}$$

Now the idea is to see what happens when $\widetilde{x} \to x$, along a given direction. So let w be a tangent vector at x, and consider a path $\varepsilon \to \widetilde{x}(\varepsilon)$, defined for $\varepsilon \in [0, \varepsilon_0)$, with initial position x and initial velocity w. (For instance, $\widetilde{x}(\varepsilon) = \exp_x(\varepsilon w)$; or in $\mathbb{R}^n$, just consider $\widetilde{x}(\varepsilon) = x + \varepsilon w$). Assume that ψ and $c(\cdot, y)$ are differentiable at x, divide both sides of (10.2) by $\varepsilon > 0$ and pass to the limit:

$$-\nabla\psi(x) \cdot w \leq \nabla_x c(x,y) \cdot w. \tag{10.3}$$

If then one changes w for $-w$, the inequality will be reversed. So necessarily

$$\nabla\psi(x) + \nabla_x c(x,y) = 0. \tag{10.4}$$

If x is given, this is an equation for y. Since our goal is to show that y is determined by x, then it will for sure help if (10.4) admits at most one solution, and this will obviously be the case if $\nabla_x c(x, \cdot)$ is *injective*. This property (injectivity of $\nabla_x c(x, \cdot)$) is in fact a classical condition in the theory of dynamical system, where it is sometimes referred to as a **twist condition**.

Three objections might immediately be raised. First, ψ is an unknown of the problem, defined by an infimum, so why would it be differentiable? Second, the injectivity of $\nabla_x c$ as a function of y seems quite hard to check on concrete examples. Third, even if c is given in the problem and a priori quite nice, why should it be differentiable at (x, y)? As a very simple example, consider the square distance function $d(x, y)^2$ on the 1-dimensional circle $S^1 = \mathbb{R}/(2\pi\mathbb{Z})$, identified with $[0, 2\pi)$:

$$d(x, y) = \min\big(|x - y|,\, 2\pi - |x - y|\big).$$

Then $d(x, y)$ is *not* differentiable as a function of x when $|y - x| = \pi$, and of course $d(x, y)^2$ is not differentiable either.

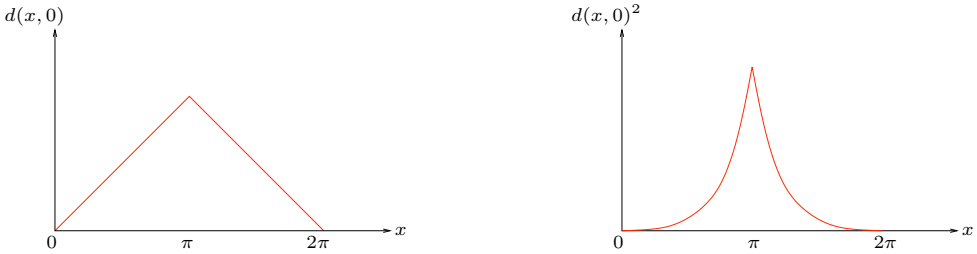


Fig. 10.1. The distance function $d(\cdot, y)$ on S^1, and its square. The upper-pointing singularity is typical. The square distance is not differentiable when $|x - y| = \pi$; still it is superdifferentiable, in a sense that is explained later.

Similar problems would occur on, say, a compact Riemannian manifold, as soon as there is no uniqueness of the geodesic joining x to y. For instance, if N and S respectively stand for the North and South Poles on S^2, then $d(x, S)$ fails to be differentiable as a function of x at $x = N$.

Of course, for almost all x this happens only for a negligible set of y's; and the cost function is differentiable everywhere else, so we might think that this is not a serious problem. But who tells us that the optimal transport will not try to take each x (or a lot of them) to a place y such that $c(x, y)$ is not differentiable??

To solve these problems, it will be useful to use some concepts from non-smooth analysis: subdifferentiability, superdifferentiability, approximate differentiability. The short answers to the above problems are that (a) under adequate assumptions on the cost function, ψ will be differentiable out of a very small set (of codimension at most 1); (b) c will be superdifferentiable because it derives from a Lagrangian, and subdifferentiable wherever ψ itself is differentiable; (c) where it exists, $\nabla_x c$ will be injective because c derives from a *strictly convex* Lagrangian.

The next three sections will be devoted to some basic reminders about differentiability and regularity in a non-smooth context. For the convenience of the non-expert reader, I shall provide complete proofs of the most basic results about these issues. Conversely, readers who feel very comfortable with these notions can skip these sections.

Differentiability and approximate differentiability

Let us start with the classical definition of differentiability:

Definition 10.1 (differentiability). *Let $U \subset \mathbb{R}^n$ be an open set. A function $f : U \to \mathbb{R}$ is said to be differentiable at $x \in U$ if there exists a vector $p \in \mathbb{R}^n$ such that*

$$f(z) = f(x) + \langle p, z - x\rangle + o(|z - x|).$$

The vector p is called the gradient of f at x, and denoted by $\nabla f(x)$; the map $w \to \langle p, w\rangle$ is the differential of f at x.

If U is an open set of a smooth Riemannian manifold M, then $f : U \to \mathbb{R}$ is said to be differentiable at x if it is so when expressed in a local chart around x; or equivalently if there is a vector $p \in T_xM$ such that

$$f(\exp_w x) = f(x) + \langle p, w\rangle + o(d(x, z)).$$

The vector p is again denoted by $\nabla f(x)$.

Differentiability is a pointwise concept, which is not invariant by, say, change of Lebesgue equivalence class: If f is differentiable or even C^∞ everywhere, by changing it on a dense countable set we may obtain a function which is discontinuous everywhere, and a fortiori not differentiable. The next notion is more flexible in this respect, since it allows for modification on a negligible set. It relies on the useful concept of **density**. Recall that a measurable set A is said to have density ρ at x if

$$\lim_{r\to 0} \frac{\mathrm{vol}\,[A \cap B_r(x)]}{\mathrm{vol}\,[B_r(x)]} = \rho.$$

It is a basic result of measure theory that *a measurable set has density 1 at almost all of its points.*

Definition 10.2 (approximate differentiability). *Let U be an open set of a Riemannian manifold M, and let $f : U \to \mathbb{R} \cup \{\pm\infty\}$ be a measurable function. Then f is said to be approximately differentiable at $x \in U$ if there is a measurable function $\widetilde{f} : U \to \mathbb{R}$, differentiable at x, and such that the set $\{\widetilde{f} = f\}$ has density 1 at x; in other words,*

$$\lim_{r\to 0} \frac{\mathrm{vol}\left[\left\{z \in B_r(x);\ f(z) = \widetilde{f}(z)\right\}\right]}{\mathrm{vol}\,[B_r(x)]} = 1.$$

Then one defines the approximate gradient of f at x by the formula

$$\widetilde{\nabla} f(x) = \nabla \widetilde{f}(x).$$

Proof that $\widetilde{\nabla} f(x)$ is well-defined. Since this concept is local and invariant by diffeomorphism, it is sufficient to treat the case when U is a subset of $\mathbb{R}^n$.

Let $\widetilde{f}_1$ and $\widetilde{f}_2$ be two measurable functions on U which are both differentiable at x and both coincide with f on a set of density 1. The problem is to show that $\nabla \widetilde{f}_1(x) = \nabla \widetilde{f}_2(x)$.

For each $r > 0$, let Z_r be the set of points in $B_r(x)$ where either $f(x) \neq \widetilde{f}_1(x)$ or $f(x) \neq \widetilde{f}_2(x)$. It is clear that $\mathrm{vol}\,[Z_r] = o(\mathrm{vol}\,[B_r(x)])$.

Since $\widetilde{f}_1$ and $\widetilde{f}_2$ are continuous at x, one can write

$$\begin{aligned}\widetilde{f}_1(x) &= \lim_{r\to 0}\ \frac{1}{\mathrm{vol}\,[B_r(x)]} \int \widetilde{f}_1(z)\, dz = \lim_{r\to 0}\ \frac{1}{\mathrm{vol}\,[B_r(x) \setminus Z_r]} \int \widetilde{f}_1(z)\, dz \\ &= \lim_{r\to 0}\ \frac{1}{\mathrm{vol}\,[B_r(x) \setminus Z_r]} \int \widetilde{f}_2(z)\, dz = \lim_{r\to 0}\ \frac{1}{\mathrm{vol}\,[B_r(x)]} \int \widetilde{f}_2(z)\, dz = \widetilde{f}_2(x).\end{aligned}$$

So let $\widetilde{f}(x)$ be the common value of $\widetilde{f}_1$ and $\widetilde{f}_2$ at x.

Next, for any $z \in Z_r$, one has

$$\widetilde{f}_1(z) = \widetilde{f}(x) + \big\langle \nabla \widetilde{f}_1(x),\, z - x \big\rangle + o(r), \qquad \widetilde{f}_2(z) = \widetilde{f}(x) + \big\langle \nabla \widetilde{f}_2(x),\, z - x \big\rangle + o(r),$$

so

$$\big\langle \nabla \widetilde{f}_1(x) - \nabla \widetilde{f}_2(x),\, z - x \big\rangle = o(r).$$

Let $w := \nabla \widetilde{f}_1(x) - \nabla \widetilde{f}_2(x)$; the previous estimate reads

$$x \in Z_r \Longrightarrow \quad \langle w, z - x \rangle = o(r). \tag{10.5}$$

If $w \neq 0$, then the set of $z \in B_r(x)$ such that $\langle w, z - x \rangle \geq r/2$ has a measure at least $K\,\mathrm{vol}\,[B_r(x)]$, for some $K > 0$. If r is small enough, then $\mathrm{vol}\,[Z_r] \leq (K/4)\mathrm{vol}\,[B_r(x)] \leq (K/2)\mathrm{vol}\,[B_r(x) \setminus Z_r]$, so

$$\mathrm{vol}\left[\left\{z \in B_r(x) \setminus Z_r;\ \langle w, z - x \rangle \geq \frac{r}{2}\right\}\right] \geq \frac{K}{2}\,\mathrm{vol}\,[B_r(x) \setminus Z_r].$$

Then (still for r small enough),

$$\frac{\displaystyle\int_{B_r(x)\setminus Z_r} \big|\langle w, z - x \rangle\big|\, dy}{\mathrm{vol}\,[B_r(x) \setminus Z_r]} \geq \frac{Kr}{4},$$

in contradiction with (10.5). The conclusion is that $w = 0$, which is what we wanted to prove. □

Regularity in a non-smooth world

Regularity is a loose concept about the control of "how fast" a function varies. In the present section I shall review some notions of regularity which apply to nonsmooth context, and act as a replacement for, say, C^1 or C^2 regularity bounds.

Definition 10.3 (Lipschitz continuity). *Let $U \subset \mathbb{R}^n$ be open, and let $f : U \to \mathbb{R}$ be given. Then*

(i) f is said to be Lipschitz if there exists $L < \infty$ such that

$$\forall x, z \in U, \qquad |f(z) - f(x)| \leq L|z - x|.$$

(ii) f is said to be locally Lipschitz if, for any $x_0 \in U$, there is a neighborhood O of x_0 in which f is Lipschitz continuous.

If U is an open subset of a Riemannian manifold M, then $f : U \to \mathbb{R}$ is said to be locally Lipschitz if it is so when expressed in local charts; or equivalently if f is Lipschitz on any compact subset of U, equipped with the geodesic distance on M.

Example 10.4. Obviously, a C^1 function is locally Lipschitz, but the converse is not true (think of $f(x) = |x|$).

Definition 10.5 (subdifferentiability and superdifferentiability). *Let U be an open set of $\mathbb{R}^n$, and $f : U \to \mathbb{R}$ a function. Then*

(i) f is said to be subdifferentiable at x, with subgradient p, if

$$f(z) \geq f(x) + \big\langle p,\, z - x \big\rangle + o(|z - x|).$$

The convex set of all subgradients p at x will be denoted by $\nabla^- f(x)$.

(ii) f is said to be uniformly subdifferentiable in U if there is a nondecreasing function $\omega : \mathbb{R}_+ \to \mathbb{R}_+$, such that $\omega(r) = o(r)$ as $r \to 0$, and

$$\forall x \in U \qquad \exists p \in \mathbb{R}^n; \qquad f(z) \geq f(x) + \big\langle p,\, z - x \big\rangle + \omega(|z - x|).$$

(iii) f is said to be locally subdifferentiable (or locally uniformly subdifferentiable) in U if each $x_0 \in U$ admits a neighborhood on which f is uniformly subdifferentiable.

If U is an open set of a smooth manifold M, and $f : U \to \mathbb{R}$ is given, then it is said to be subdifferentiable at some point x, or uniformly subdifferentiable in U, if it is so when expressed in local charts.

Corresponding notions of superdifferentiability and supergradients are obtained in an obvious way by just reversing the signs of the inequalities. The convex set of supergradients for f at x is denoted by $\nabla^+ f(x)$.

Examples 10.6. If f is minimum at $x_0 \in U$, then 0 is a subgradient of f at x_0, whatever the regularity of f. If f has a subgradient p at x and g is smooth, then $f + g$ has a subgradient $p + \nabla g(x)$ at x. If f is convex in U, then it is (uniformly) subdifferentiable at every point in U, by the well-known inequality

$$f(z) \geq f(x) + \big\langle p,\, z - x \big\rangle,$$

which holds true as soon as $p \in \partial f(x)$ and $[x, y] \subset U$. If f is the sum of a convex function and a smooth function, then it is also uniformly subdifferentiable.

It is obvious that differentiability implies both subdifferentiability and superdifferentiability. The converse is true, as shown by the next statement.

Proposition 10.7 (subdifferentiability and superdifferentiability imply differentiability). *Let U be an open set of a smooth Riemannian manifold M, and let $f : U \to \mathbb{R}$ be a function. Then f is differentiable at x if and only if it is both subdifferentiable and superdifferentiable there; and then*

$$\nabla^- f(x) = \nabla^+ f(x) = \{\nabla f(x)\}.$$

Proof of Proposition 10.7. The only nontrivial implication is that if f is both subdifferentiable and superdifferentiable, then it is differentiable. Since this statement is local and invariant by diffemorphism, let us pretend that $U \subset \mathbb{R}^n$. So let $p \in \nabla^- f(x)$ and $q \in \nabla^+ f(x)$; then

$$f(z) - f(x) \geq \big\langle p,\, z - x \big\rangle - o(|z - x|);$$
$$f(z) - f(x) \leq \big\langle q,\, z - x \big\rangle + o(|z - x|).$$

It follows that $\langle p - q, z - x \rangle \leq o(|z - x|)$, which means

$$\lim_{z \to x;\ z \neq x} \left\langle p - q,\, \frac{z - x}{|z - x|} \right\rangle = 0.$$

Since the unit vector $(z - x)/|z - x|$ can take arbitrary fixed values in the unit sphere as $z \to x$, it follows that $p = q$. Then

$$f(z) - f(x) = \big\langle p,\, z - x \big\rangle + o(|z - x|),$$

which means that f is indeed differentiable at x. This also shows that $p = q = \nabla f(x)$, and the proof is complete. □

The next proposition summarizes some of the most important results about the links between regularity and differentiability:

Theorem 10.8. *Let U be an open subset of a smooth Riemannian manifold M, and let $f : U \to \mathbb{R}$ be a function. Let n be the dimension of M. Then*

(i) If f is continuous, then it is subdifferentiable on a dense subset of U, and also superdifferentiable on a dense subset of U;

(ii) If f is locally Lipschitz, then it is differentiable almost everywhere (with respect to the volume measure);

(iii) If f is locally subdifferentiable (or locally superdifferentiable), then it is differentiable out of a countably $(n-1)$-rectifiable set.

Remark 10.9. Statement (ii) is known as **Rademacher's theorem**. The conclusion in Statement (iii) is stronger than differentiability almost everywhere, since a $(n-1)$-rectifiable set has dimension $n-1$, and is therefore negligible. In fact, as we shall see very soon, the local subdifferentiability property is stronger than the local Lipschitz property. Reminders about the notion of countable rectifiability are provided in the Appendix.

Proof of Theorem 10.8. First we can cover U by a countable collection of small open sets U_k, each of which is diffeomorphic to an open subset O_k of $\mathbb{R}^n$. Then, since all the concepts involved are local and invariant under diffeomorphism, we may work in O_k. So in the sequel, I shall pretend that U is a subset of $\mathbb{R}^n$.

Let us start with the proof of (i). Let f be continuous on U, and let V be an open subset of U; the problem is to show that f admits at least one point of subdifferentiability in V. So let $x_0 \in V$, and let $r > 0$ be so small that $\overline{B(x_0, r)} \subset V$. Let g be defined on B by $g(x) := f(x) + |x - x_0|^2/2\varepsilon$. Since f is continuous, g attains its minimum on $\overline{B}$. But g on ∂B it is bounded below by $r^2/\varepsilon - M$, where M is an upper bound for $|f|$ on $\overline{B}$. If $\varepsilon < r^2/(2M)$, then $g(x_0) \le f(x_0) \le M < r^2/\varepsilon - M < \inf_{\partial B} g$; so g cannot achieve its minimum on ∂B, and has to achieve it at some point $x_1 \in B$. Then g is subdifferentiable at x_1, and therefore f also. This establishes Statement (i).

The other two statements are more tricky. Let us start the proof of (ii). Let $f : U \to \mathbb{R}$ be a Lipschitz function. For $v \in \mathbb{R}^n$ and $x \in U$, define

$$D_v f(x) := \lim_{t \to 0} \left[\frac{f(x + tv) - f(x)}{t} \right], \tag{10.6}$$

provided that this limit exists. The problem is to show that for almost all x, there is a vector $p(x)$ such that $D_v f(x) = \langle p(x), v \rangle$ and the limit in (10.6) is uniform in, say, $v \in S^{n-1}$. Since the functions $[f(x + tv) - f(x)]/t$ are uniformly Lipschitz in v, it is actually enough to prove the pointwise convergence (that is, the mere existence of $D_v f(x)$), and then the limit will be automatically uniform. So the goal is to show that for almost all x, the limit $D_v f(x)$ exists for all v, and is linear in v.

It is easily checked that

(a) $D_v f(x)$ is homogeneous in v: $D_{tv} f(x) = t\, D_v f(x)$;

(b) $D_v f(x)$ is a Lipschitz function of v on its domain: in fact, $|D_v f(x) - D_w f(x)| \le L\, |v - w|$, where $L = \|f\|_{\mathrm{Lip}}$;

(c) If $D_w f(x) \to \ell$ as $w \to v$, then $D_v f(x) = \ell$; this comes from the estimate

$$\sup_t \left| \left(\frac{f(x + tv) - f(x)}{t} \right) - \left(\frac{f(x + tv_k) - f(x)}{t} \right) \right| \le \|f\|_{\mathrm{Lip}}\, |v - v_k|.$$

For each $v \in \mathbb{R}^n$, let A_v be the set of $x \in \mathbb{R}^n$ such that $D_v f(x)$ does not exist. The first claim is that *each A_v has zero Lebesgue measure.* This is obvious if $v \neq 0$. Otherwise, let $H = v^\perp$ be the hyperplane orthogonal to v, passing through the origin. For each $x_0 \in H$, let $L_{x_0} = x_0 + \mathbb{R}v$ be the line parallel to v, passing through x_0. The nonexistence of $D_v f(x)$ at $x = x_0 + t_0 v$ is equivalent to the nondifferentiability of $t \to f(x + tv)$ at $t = t_0$. Since $t \to f(x+tv)$ is Lipschitz $\mathbb{R} \to \mathbb{R}$, it follows from a well-known result of real analysis that it is differentiable for λ_1-almost all $t \in \mathbb{R}$, where λ_1 stands for the one-dimensional Lebesgue measure. So $\lambda_1[A_v \cap L_{x_0}] = 0$. Then by Fubini's theorem, $\lambda_n[A_v] = \int_H \lambda_1[A_v \cap L_{x_0}]\, dx_0 = 0$, where λ_n is the n-dimensional Lebesgue measure, and this proves the claim.

Now, the problem consists in extending the function $D_v f$ in such a way that it is a *linear* (not just homogeneous) function of v. Let $v \in \mathbb{R}^n$, and let ζ be a smooth compactly supported function. Then, by the dominated convergence theorem,

$$\begin{aligned}(\zeta * D_v f)(x) &= \int \zeta(x-y) \lim_{t\to 0} \left[\frac{f(y+tv) - f(y)}{t}\right] dy \\ &= \lim_{t\to 0} \frac{1}{t} \int \zeta(x-y)\, [f(y+tv) - f(y)]\, dy \\ &= \lim_{t\to 0} \frac{1}{t} \int \zeta(x-y)\, [\zeta(x-y-tv) - \zeta(x-y)]\, f(y)\, dy \\ &= \int \langle \nabla\zeta(x-y), v\rangle\, f(y)\, dy.\end{aligned}$$

(Note that $\zeta * D_v f$ is well-defined for any x, even if $D_v f$ is defined only for almost all x.) So $\zeta * D_v f$ depends *linearly* on v. In particular, if v and w are any two vectors in $\mathbb{R}^n$, then

$$\zeta * [D_{v+w} f - D_v f - D_w f] = 0.$$

Since ζ is arbitrary, it follows that

$$D_v f(x) + D_w f(x) = D_{v+w} f(x) \tag{10.7}$$

for almost all $x \in \mathbb{R}^n \setminus (A_v \cap A_w \cap A_{v+w})$, that is, for almost all $x \in \mathbb{R}^n$.

Now it is easy to conclude. Let $B_{v,w}$ be the set of all $x \in \mathbb{R}^n$ such that $D_v f(x)$, $D_w f(x)$ or $D_{v+w} f(x)$ is not well-defined, or (10.7) does not hold true. Let $(v_k)_{k\in\mathbb{N}}$ be a dense sequence in R^n, and let $B := \bigcup_{j,k\in\mathbb{N}} B_{v_j, v_k}$. Then B is still Lebesgue-negligible, and for each $x \notin B$ we have

$$D_{v_j + v_k} f(x) = D_{v_j} f(x) + D_{v_k} f(x). \tag{10.8}$$

Since $D_v f(x)$ is a Lipschitz continuous function of v, it can be extended uniquely into a Lipschitz continuous function, defined for all $x \notin B$ and $v \in \mathbb{R}^n$, which turns out to be $D_v f(x)$ in view of Property (c). By passing to the limit in (10.8), we see that $D_v f(x)$ is an additive function of v. We already know that it is a homogeneous function of v, so it is in fact linear. This concludes the proof of (ii).

Now let us turn to the proof of (iii). Before going on, I shall first explain in an informal way the **main idea of the proof of statement (iii)**. Suppose for simplicity that we are dealing with a convex function in $\mathbb{R}^n$. If p lies in the subdifferential $\partial\psi(x)$ of ψ at x, then for all $z \in \mathbb{R}^n$,

$$\psi(z) \geq \psi(x) + \big\langle p,\, z - x\big\rangle.$$

In particular, if $p \in \partial\psi(x)$ and $p' \in \partial\psi(x')$, then

$$\big\langle p - p',\ x - x' \big\rangle \geq 0.$$

If ψ is not differentiable at x, this means that $\partial\psi(x)$ is not reduced to a single element, so it should contain a line segment $[p, p'] \subset \mathbb{R}^n$. For these heuristic explanations, let us *fix* p and p', and consider the set Σ of all $x \in \mathbb{R}^n$ such that $[p, p'] \subset \partial\psi(x)$. Then $\langle p - p', x - x' \rangle \geq 0$ for all $x, x' \in \Sigma$. By exchanging the roles of p and p', we see that actually $\langle p - p', x - x' \rangle = 0$. This implies that Σ is included in a single hyerplane, orthogonal to $p - p'$; in particular its dimension is at most $n - 1$.

Now comes the rigorous argument. Let Σ be the set of points x such that $\nabla^- \psi(x)$ is not reduced to a single element. Since $\nabla^- \psi(x)$ is a convex set, for each $x \in \Sigma$ there is a nontrivial segment $[p, p'] \subset \nabla^- \psi(x)$. So

$$\Sigma = \bigcup_{\ell \in \mathbb{N}} \Sigma^{(\ell)},$$

where $\Sigma^{(\ell)}$ is the set of points x such that $\nabla^- \psi(x)$ contains a segment $[p, p']$ of length $1/\ell$. To conclude, it is sufficient to show that each $\Sigma^{(\ell)}$ is countably $(n-1)$-rectifiable, and for that it is sufficient to show that for each $x \in \Sigma^{(\ell)}$ the dimension of the tangent cone $T_x \Sigma^{(\ell)}$ is at most $n - 1$. (In case of need, see the Appendix for reminders about the notions of countable rectifiability and tangent cone.)

So let $x \in \Sigma^{(\ell)}$, and let $q \in T_x \Sigma^{(\ell)}$, $q \neq 0$. By assumption, there is a sequence $x_k \in \Sigma^{(\ell)}$ such that

$$\frac{x_k - x}{t_k} \longrightarrow q.$$

In particular $|x - x_k|/t_k$ converges to the finite, nonzero limit $|q|$.

Now, for all k, there is a segment $[p_k, p_k']$, of length ℓ^{-1}, that is contained in $\nabla^- \psi(x_k)$. By compactness, up to extraction of a subsequence one has $(x_k, p_k) \to (x, p)$ and $(x_k, p_k') \to (x, p')$. By continuity of $\nabla^- \psi$, both p and p' belong to $\nabla^- \psi(x)$. Then the two inequalities

$$\begin{cases} \psi(x) \geq \psi(x_k) + \big\langle p_k',\ x - x_k \big\rangle - \omega(|x - x_k|) \\ \\ \psi(x_k) \geq \psi(x) + \big\langle p,\ x - x_k \big\rangle - \omega(|x - x_k|) \end{cases}$$

combine to yield

$$\big\langle p - p_k',\ x - x_k \big\rangle \geq -2\,\omega(|x - x_k|).$$

So

$$\Big\langle p - p_k',\ \frac{x - x_k}{t_k} \Big\rangle \geq -2\ \frac{\omega(|x - x_k|)}{|x - x_k|}\ \frac{|x - x_k|}{t_k}.$$

Passing to the limit, we find

$$\langle p - p', q \rangle \geq 0.$$

But the roles of p and p' can be exchanged, so actually

$$\langle p - p', q \rangle = 0.$$

Since $p - p'$ is nonzero, this means that q belongs to the hyperplane $(p - p')^\perp$. So for each $x \in \Sigma^{(\ell)}$, the tangent cone $T_x \Sigma^{(\ell)}$ is included in a hyperplane. This concludes the proof. □

Semi-convexity and semi-concavity

Convexity can be expressed without any reference to smoothness, yet it implies a lower bound on the Hessian. More generally, in nonsmooth analysis, convexity-type estimates are often used as a replacement for second-order derivative bounds. In this respect the notion of semi-convexity is extremely convenient.

Definition 10.10 (semi-convexity). *Let U be an open set of a smooth Riemanian manifold and let $\omega : \mathbb{R}_+ \to \mathbb{R}_+$ be continuous nondecreasing, such that $\omega(r) = o(r)$ as $r \to 0$. A function $f : U \to \mathbb{R} \cup \{+\infty\}$ is said to be semi-convex with modulus ω if, for any constant-speed geodesic path $(\gamma_t)_{0 \le t \le 1}$, whose image is included in U,*

$$f(\gamma_t) \le (1-t) f(\gamma_0) + t f(\gamma_1) + t(1-t)\, \omega\big(d(\gamma_0, \gamma_1)\big). \tag{10.9}$$

It is said to be locally semi-convex if for each $x_0 \in U$ there is a neighborhood V of x_0 in U such that (10.9) holds true as soon as $\gamma_0, \gamma_1 \in V$; or equivalently if (10.9) holds true for some modulus ω_K as long as γ stays in a compact subset K of U.

Similar definitions for semi-concavity and local semi-concavity are obtained in an obvious way by reversing the sign of the inequality in (10.9).

Example 10.11. In $\mathbb{R}^n$, semi-convexity with modulus ω is equivalent to the requirement that for all $x, y \in \mathbb{R}^n$ and $t \in [0,1]$,

$$f\big((1-t)x + ty\big) \le (1-t)\, f(x) + t\, f(y) + t(1-t)\, \omega(|x-y|).$$

In particular, when $\omega = 0$ this is the usual notion of convexity. In the case $\omega(r) = Cr^2/2$, there is a differential characterization of semi-convexity in terms of Hessian matrices: $f : \mathbb{R}^n \to \mathbb{R}$ is semi-convex with modulus $\omega(r) = Cr^2/2$ if and only if $\nabla^2 f \ge -C I_n$. (If f is not twice differentiable, then $\nabla^2 f$ should be interpreted as the distributional gradient.)

A well-known theorem of convex analysis states that a convex function is subdifferentiable everywhere in the interior of its domain. The next Proposition generalizes this statement to semi-convex functions.

Proposition 10.12 (Semi-convexity and subdifferentiability).
Let M be a smooth Riemannian manifold. Then

(i) If $\psi : M \to \mathbb{R} \cup \{+\infty\}$ is locally semi-convex, then it is locally subdifferentiable in the interior of its domain $D := \psi^{-1}(\mathbb{R})$; and ∂D is countably $(n-1)$-rectifiable;

(ii) Conversely, if U is an open subset of M, and $\psi : U \to \mathbb{R}$ is locally subdifferentiable, then it is also locally semi-convex.

Similar statements hold true with "subdifferentiable" replaced by "superdifferentiable" and "semi-convex" replaced by "semi-concave".

Remark 10.13. This proposition implies that *local* semi-convexity and local subdifferentiability are basically the same. But semi-convexity can also be stated in a *global* version.

Proof of Proposition 10.12. First, we can cover M by a countable union of open sets U, each of which satisfies the following property: There is a function $\omega(r) = o(r)$ such that for any geodesic γ with endpoints $\gamma_0, \gamma_1 \in U$,

$$\psi(\gamma_t) \le (1-t)\, \psi(\gamma_0) + t\, \psi(\gamma_1) + t(1-t)\, \omega\big(d(\gamma_0, \gamma_1)\big), \tag{10.10}$$

where $\omega(r) = o(r)$. To prove the statement, it is sufficient to work in U.

Inequality (10.10) implies that the domain D of ψ is a geodesically convex subset of U; more precisely, if $x_0, x_1 \in D \cap U$, and γ is a geodesic joining x_0 to x_1, then the image of γ is entirely contained in D. It is not difficult to deduce that for any $x \in \overline{D}$, the tangent cone $T_x D$ is convex. This leaves two possibilities: Either $T_x D$ is included in a half-space, or it is the whole of $T_x M$.

At this point I shall admit a few facts which can be deduced from a bit of work in Riemannian geometry. Assume that $T_x D = T_x M$. This means that for any tangent vectors $v_1, \ldots, v_N$ in $T_x M$, one can find N points $y_1, \ldots, y_N$ in the neighborhood of x, in such a way that $y_i - x$ is very close to v_i, for each $i \in \{1, \ldots, N\}$. (By abuse of notation, I write $y_i - x$ for the initial velocity of the unique geodesic joining x to y_i.) If the vectors v_i are well chosen, then the vectors $y_i - x$ will form an affine basis of $T_x M$, meaning that any tangent vector $v \in T_x M$ can be written as an affine combination of the vectors $y_i - x$. Then it can be checked that there is a neighborhood O of x such that any point in O can be obtained by taking successive barycenters of the points $y_1, \ldots, y_N$. (By assumption the barycenter $[y, y']_t$ of y and y' with coefficient t is the point γ_t, where γ is a geodesic with $\gamma_0 = y$ and $\gamma_1 = y'$.) So ψ will be bounded above, on O by $\max(\psi(y_1), \ldots, \psi(y_N))$, plus a remainder which only depends on the neighborhood and the modulus of semi-convexity.

To summarize: If $T_x D = T_x M$, then ψ is bounded above in a neighborhood of x; in particular, x lies in the interior of D. This shows that for each $x \in \partial D$, $T_x D$ is included in a half-space, and so, by a theorem recalled in the Appendix, ∂D is countably $(n-1)$-rectifiable. This also shows that ψ is *locally bounded above* in the interior of its domain.

Next let again $x \in D$. If W is a sufficiently small neighborhood of x, then for any point $y \in W$ there exists $\widetilde{y} \in W$ such that x is a midpoint between y and $\widetilde{y}$, so

$$\psi(x) \le \frac{1}{2}[\psi(y) + \psi(\widetilde{y})].$$

Since $\psi(\widetilde{y})$ is bounded above and $\psi(x)$ is a fixed number, it follows that $\psi(y)$ is bounded below for $y \in W$. The conclusion is that ψ is not only locally bounded above, but also locally bounded below in the interior of its domain.

Now let us show that ψ is actually *locally Lipschitz*. Let again $x \in D$; there is a neighborhood V of x such that whenever y and y' belong to V, and $\lambda \in [0,1]$ is arbitrary, then there is y_λ such that $y' = [y, y_\lambda]_\lambda$. Then

$$\psi(y') \le (1-\lambda)\,\psi(y) + \lambda\,\psi(y_\lambda) + \lambda(1-\lambda)\,\omega\big(d(y, y_\lambda)\big),$$

so

$$\frac{\psi(y') - \psi(y)}{d(y,y')} = \frac{\psi(y') - \psi(y)}{\lambda\, d(y, y_\lambda)} \le \frac{\psi(y_\lambda) - \psi(y)}{d(y, y_\lambda)} + \frac{\omega(d(y, y_\lambda))}{d(y, y')}.$$

Now choose $\lambda = r\, d(y, y')$, so that $d(y, y_\lambda) = r$ (this is possible if $r > 0$ is small enough), and then

$$\frac{\psi(y') - \psi(y)}{d(y,y')} \le \frac{2M}{r} + \frac{\omega(r)}{r},$$

where M is an upper bound for $|\psi|$. So the ratio $(\psi(y') - \psi(y))/d(y, y')$ is uniformly bounded above for y, y' in V. By symmetry (exchange y and y'), there is also a uniform lower bound for that ratio, and in the end ψ is indeed Lipschitz in V.

The next claim is that $\nabla^- \psi$ *is continuous.* This means that if sequences $x_k \in \Omega$ and $p_k \in T_{x_k} M$ are given, with $p_k \in \nabla^- \psi(x_k)$ and $(x_k, p_k) \to (x, p)$, then $p \in \nabla^- \psi(x)$.

Indeed, let z be arbitrary in Ω, and for each $k \in \mathbb{N}$, let $(\gamma_k(t))_{0 \le t \le 1}$ be a constant speed geodesic joining x_k to z. Since $(\gamma_k)_{k \in \mathbb{N}}$ is a bounded family of geodesics (in particular,

uniformly Lipschitz), by Ascoli's theorem it converges uniformly, up to extraction, and it is not hard to show that the limit has to be a geodesic curve γ joining x to y. By localization, one may assume that this geodesic curve is unique; and then one can show that $w_k := \dot\gamma_k(0)$ converges to $w := \dot\gamma(0)$.

Assume that not only ψ is subdifferentiable at x_k with subgradient p_k, but in addition there is a uniform modulus of subdifferentiability:

$$\psi(z) \geq \psi(x_k) + \langle p_k, w_k\rangle - \omega\big(d(x_k,z)\big), \tag{10.11}$$

then this inequality passes to the limit as $(x_k,p_k)\to(x,p)$ and $w_k\to w$, and this implies that p is a subgradient of ψ at x.

So it is sufficient to establish (10.11). For simplicity, I shall drop the index k. Let ψ be semi-convex with modulus ω, and subdifferentiable at x, with subgradient p; let further y be an arbitrary point in the neighborhood of x, let γ be a geodesic joining x to y, and $w := \dot\gamma(0)$. Then

$$\psi(\gamma(t)) \leq (1-t)\,\psi(x) + t\psi(y) + t(1-t)\,\omega(|w|),$$

so

$$\frac{\psi(\gamma(t))-\psi(x)}{t|w|} \leq \frac{\psi(y)-\psi(x)}{|w|} + (1-t)\,\frac{\omega(|w|)}{|w|}.$$

On the other hand, by subdifferentiability,

$$\frac{\psi(\gamma(t))-\psi(x)}{t|w|} \geq \frac{\langle p, tw\rangle}{t|w|} - \frac{o(t|w|)}{t|w|} = \Big\langle p, \frac{w}{|w|}\Big\rangle - \frac{o(t|w|)}{t|w|}.$$

The combination of both inequalities yields

$$\frac{\langle p,w\rangle}{|w|} - \frac{o(t|w|)}{t|w|} \leq \frac{\psi(y)-\psi(x)}{|w|} + (1-t)\,\frac{\omega(|w|)}{|w|}.$$

If $t\to 0$, this implies

$$\frac{\langle p,w\rangle}{|w|} \leq \frac{\psi(y)-\psi(x)}{|w|} + \frac{\omega(|w|)}{|w|}.$$

But this is exactly the desired inequality (10.11).

Next, I claim that *the subdifferential of ψ is locally bounded.* Indeed, let x be a point of subdifferentiability of ψ, and let y be an arbitrary point in the neighborhood of x; let γ be a geodesic joining x to y, and $w := \dot\gamma(0)$, then the subdifferentiability implies

$$\Big\langle p, \frac{w}{|w|}\Big\rangle \leq \frac{\psi(y)-\psi(x)}{d(x,y)} + \frac{o(d(x,y))}{d(x,y)}.$$

It is possible to let $y\to x$ while keeping $w/|w| \simeq \sigma$, where σ is an arbitrary unit vector in T_xM. Then

$$\langle p,\sigma\rangle \leq \limsup_{y\to x}\ \frac{\psi(y)-\psi(x)}{d(x,y)}.$$

In particular,

$$|p| \leq \limsup_{y\to x}\ \frac{\psi(y)-\psi(x)}{d(x,y)},$$

which is bounded above by a finite constant in view of Step 4.

Now, let $x\in\Omega$. By Theorem 10.8(i), there is a sequence $x_k\to x$ with $x_k\in\Omega$ and $\nabla^-\psi(x_k)\neq\emptyset$. So let $p_k\in\nabla^-\psi(x_k)$. By Step 6, the family p_k is bounded as $k\to\infty$; so it is

possible to extract a subsequence of (x_k, p_k) which converges to some $(x, p) \in T_x M$. Then by Step 5, $p \in \nabla^- \psi(x)$. The conclusion is that ψ is subdifferentiable at x. In fact, it is even uniformly subdifferentiable in a neighborhood of x, with modulus ω. This concludes the proof of Statement (i).

Statement (ii) is much easier. Let $x \in U$, and let V be a small neighborhood of x, such that f is uniformly subdifferentiable in V with modulus ω. Without loss of generality, assume that $\omega(r)/r$ is a nonincreasing function of r. Let $W \subset V$ be a neighborhood of x, small enough that any two points y, y' in W can be joined by a unique geodesic $\gamma^{y,y'}$, whose image is contained in V; by abuse of notation I shall write $y' - y$ for the initial velocity of $\gamma^{y,y'}$.

Let then γ be a geodesic such that $\gamma_0, \gamma_1 \in V$; let $t \in [0,1]$, and let $p \in \nabla^- f(\gamma_t)$. It follows from the subdifferentiability that

$$f(\gamma_1) \leq f(\gamma_t) + \langle p,\, \gamma_1 - \gamma_t \rangle + \omega\big(d(\gamma_t, \gamma_1)\big).$$

Since $d(\gamma_t, \gamma_1) = (1-t)\, d(\gamma_0, \gamma_1)$ and $\omega(r)/r$ is nonincreasing, it follows that

$$f(\gamma_1) \leq f(\gamma_t) + \langle p,\, \gamma_1 - \gamma_t \rangle + (1-t)\, \omega\big(d(\gamma_0, \gamma_1)\big). \tag{10.12}$$

Similarly,

$$f(\gamma_0) \leq f(\gamma_t) + \langle p,\, \gamma_0 - \gamma_t \rangle + t\, \omega\big(d(\gamma_0, \gamma_1)\big). \tag{10.13}$$

Now take the linear combination of (10.12) and (10.13) with coefficients t and $1-t$: Since $t(\gamma_1 - \gamma_t) + (1-t)(\gamma_0 - \gamma_t) = 0$ (in $T_{\gamma_t} M$), we recover

$$(1-t)\, f(\gamma_0) + t\, f(\gamma_1) - f(\gamma_t) \leq 2\, t\, (1-t)\, \omega(d(\gamma_0, \gamma_1)).$$

This proves that f is semi-convex in W. □

Assumptions on the cost function

Let M be a Riemannian manifold, let $\mathcal{X}$ be a closed subset of M, let $\mathcal{Y}$ be an arbitrary Polish space, and let $c : M \times \mathcal{Y} \to \mathbb{R}$ be a continuous cost function. (Most of the time, we shall have $\mathcal{X} = M = \mathcal{Y}$.) We shall impose certain assumptions on the behavior of c as a function of x, when x varies in the interior (in M) of $\mathcal{X}$. They will be chosen among the following list:

(Super) *$c(x, y)$ is everywhere superdifferentiable as a function of x.*

(Twist) *Where it exists, $\nabla_x c(x, \cdot)$ is injective: $\nabla_x c(x, y) = \nabla_x c(x, y') \Longrightarrow y = y'$.*

(Lip) *$c(x, y)$ is locally Lipschitz as a function of x, uniformly in y.*

(SC) *$c(x, y)$ is locally semi-concave as a function of x, uniformly in y.*

(locLip) *$c(x, y)$ is locally Lipschitz as a function of x, locally in y.*

(locSC) *$c(x, y)$ is locally semi-concave as a function of x, locally in y.*

$(\mathbf{H}\infty)_1$ For any x and for any measurable set S which does not "lie on one side of x", in the sense that $T_x S$ is not contained in a half-space, there is a finite collection of elements $z_1, \ldots, z_k \in S$, and a small open ball B containing x, such that for any y outside of a compact set,

$$\inf_{w \in B} c(w, y) \geq \inf_{1 \leq j \leq k} c(z_j, y).$$

(H∞)$_2$ For any x and any neighborhood U of x there is a small ball B containing x such that

$$\lim_{y\to\infty} \sup_{w\in B} \inf_{z\in U} [c(z,y) - c(w,y)] = -\infty.$$

Our theorems of solvability of the Monge problem will be expressed in terms of these assumptions. Before going any further, I shall give some informal explanations about **(H∞)$_1$** and **(H∞)$_2$**, which probably look obscure to the reader. Both of them are assumptions about the behavior of $c(x,y)$ as $y \to \infty$, and therefore they are void if y varies in a compact set. They are essentially quantified versions of the following statement: For any y it is possible to lower the cost to go from x to y, by starting from a well-chosen nearby point z. For instance, if c is a radially symmetric cost on $\mathbb{R}^n \times \mathbb{R}^n$, then I would choose z very close to x, "opposite to y".

In the rest of this section, I shall discuss some simple sufficient conditions for all these assumptions to hold true. The first result is that Conditions **(Super)**, **(Twist)**, **(locLip)** and **(locSC)** are satisfied by many Lagrangian cost functions.

Proposition 10.14 (Properties of Lagrangian cost functions). *On a smooth Riemannian manifold M, let $c(x,y)$ be a cost function associated with a C^1 Lagrangian $L(x,v,t)$. Assume that any $x, y \in M$ can be joined by at least one C^1 minimizing curve. Then*

(i) For any $(x,y) \in M \times M$, and any C^1 action-minimizing curve γ connecting x to y, the tangent vector $-\nabla_v L(x, \dot\gamma_0, 0) \in T_x M$ is a supergradient for $c(\cdot, y)$ at x; in particular, c is superdifferentiable at (x,y) as a function of x.

(ii) If L is strictly convex as a function of v, and minimizing curves are uniquely determined by their initial position and velocity, then c satisfies a twist condition: If c is differentiable at (x,y) as a function of x, then y is uniquely determined by x and $\nabla_x c(x,y)$. Moreover,

$$\nabla_x c(x,y) + \nabla_v L(x, \dot\gamma(0), 0) = 0,$$

where γ is the unique minimizing curve joining x to y.

(iii) If L has the property that for any two compact sets K_0 and K_1, the velocities of minimizing curves starting in K_0 and ending in K_1 are uniformly bounded, then then c is locally Lipschitz and locally semi-concave as a function of x, locally in y.

Example 10.15. Consider the case $L(x,v,t) = |v|^2$. Then $\nabla_v L = 2v$; and (i) says that $-2v_0$ is a supergradient of $d(\cdot, y)^2$ at x, where v_0 is the velocity used to go from x to y. This is a generalization of the usual formula in Euclidean space:

$$\nabla_x(|x-y|^2) = 2(x-y) = -2(y-x).$$

Also (ii) says that this cost function satisfies the twist property.

Remark 10.16. The requirements in (ii) and (iii) are fulfilled if the Lagrangian L is C^2 and strictly convex superlinear as a function of v (Recall Example 7.5). But it also holds true for other interesting cases such as $L(x,v,t) = |v|^{1+\alpha}$, $0 < \alpha < 1$.

Remark 10.17. Part (i) of Proposition 10.14 means that the behavior of the (square) distance function is typical: if one plots $c(x,y)$ as a function of x, for fixed y, one will always see *upper-pointing* crests as in Figure 10.1, never downward-pointing ones.

Proof of Proposition 10.14. This Proposition is mainly based on the formula of first variation. Let (x, y) be given, and let $\gamma(t)_{0\le t\le 1}$ be a minimizing curve, C^1 as a function of t, joining x to y. Let $\widetilde{\gamma}$ be another curve, *not necessarily minimizing*, joining $\widetilde{x}$ to $\widetilde{y}$. Assume that $\widetilde{x}$ is very close to x, so that there is a unique geodesic joining x to $\widetilde{x}$; by abuse of notation, I shall write $\widetilde{x} - x$ for the initial velocity of this geodesic. Similarly, let us assume that $\widetilde{y}$ is very close to y. Then, by the formula of first variation,

$$\mathcal{A}(\widetilde{\gamma}) = \int_0^1 L\big(\dot\gamma(t), \gamma(t), t\big)\, dt + \Big[\nabla_v L\big(\gamma(1), \dot\gamma(1), 1\big) \cdot (\widetilde{y} - y) - \nabla_v L\big(\gamma(0), \dot\gamma(0), 0\big) \cdot (\widetilde{x} - x)\Big] + \omega(|\widetilde{x} - x| + |\widetilde{y} - y|), \quad (10.14)$$

where $\omega(r)/r \to 0$, and ω only depends on the behavior of the manifold in a neighborhood of γ, and on a modulus of continuity for the derivatives of L, on a neighborhood of $\{(\gamma(t), \dot\gamma(t), t)_{0\le t\le 1}\}$. Without loss of generality, we may assume that $\omega(r)/r$ is nonincreasing.

From this formula the proof of Statement (i) follows immediately: Let $\widetilde{\gamma}$ be a curve joining $\widetilde{x}$ to $\widetilde{y} = y$ (this curve can be constructed smoothly in local charts), then $c(\widetilde{x}, y) \le \mathcal{A}(\widetilde{\gamma})$, so (10.14) gives

$$c(\widetilde{x}, y) \le c(x, y) - \langle \nabla_v L(x, v, 0)\rangle + \omega(|\widetilde{x} - x|),$$

which shows that $-\nabla_v L(x, v, 0)$ is a supergradient of $c(\cdot, y)$ at x.

Now for the proof of (ii): If $c(\cdot, y)$ is not only superdifferentiable but plainly differentiable, then by Proposition 10.7 there is just one supergradient, which is the gradient, so $-\nabla_v L(x, v, 0) = \nabla_x c(x, y)$. Since L is strictly convex with respect to the v variable, this equation determines v uniquely. By assumption, this in turn determines the whole geodesic γ, and in particular y.

Finally, we turn to Statement (iii). When we let x and y vary in small balls, the velocity v along the minimizing curves will be bounded by assumption; so the function ω will also be uniform. Then $c(x, y)$ is locally superdifferentiable as a function of x, and the conclusion follows from Proposition 10.12. □

Proposition 10.14 is basically all that is needed to treat quite general cost functions on a compact Riemannian manifold. But for noncompact manifolds, it might be very difficult to check Assumptions **(Lip)**, **(SC)** or **(H∞)**. Here are a few examples where this can be done.

Example 10.18. Gangbo and McCann have considered cost functions of the form $c(x, y) = c(x - y)$ on $\mathbb{R}^n \times \mathbb{R}^n$, satisfying the following assumption: *For any given $r > 0$ and $\theta \in (0, \pi)$, if $|y|$ is large enough then there is a cone $K_{r,\theta}(y, e)$, with apex y, direction e, height h and angle θ, such that c takes its maximum on $K_{r,\theta}(y, e)$ at y.* Let us check briefly that this assumption implies $\mathbf{(H\infty)}_1$. (The reader who feels that both assumptions are equally obscure may very well skip this and jump directly to Example 10.19.) Let x and S be given such that $T_x S$ is included in no half-space. So for each direction $e \in S^{n-1}$ there are points z_+ and z_- in S, each of which lies on one side of the hyperplane passing through z and having direction e. By a compactness argument, one can find a finite collection of points $z_1, \ldots, z_k$ in S, an angle $\theta < \pi$ and a positive number $r > 0$ such that for all $e \in S^{n-1}$ and for any w close enough to x, the truncated cone $K_{r,\theta}(w, e)$ contains at least one of the z_j. Equivalently, $K_{r,\theta}(w - y, e)$ contains $z_j - y$. But by assumption, for $|w - y|$ large enough there is a cone $K_{r,\theta}(w - y, e)$ such that $c(z - y) \le c(w - y)$ for all $z \in K_{r,\theta}(w - y, e)$. This inequality applies to $z = z_j$ (for some j), and then $c(z_j - y) \le c(w - y)$.

Example 10.19. As a particular case of the previous example, Condition $(\mathbf{H}\infty)_1$ holds true if $c = c(x - y)$ is radially symmetric and strictly increasing as a function of $|x - y|$.

Example 10.20. Gangbo and McCann considered cost functions that also satisfied $c(x, y) = c(x - y)$ with c convex and superlinear. This assumption implies $(\mathbf{H}\infty)_2$. Indeed, if x in $\mathbb{R}^n$ and $\varepsilon > 0$ are given, let $z = x - \varepsilon(x - y)/|x - y|$; it suffices to show that

$$c(z - y) - c(x - y) \xrightarrow[y\to\infty]{} -\infty,$$

or equivalently, with $h = z - y$,

$$c(h) - c\Big(h\big(1 - \frac{\varepsilon}{|h|}\big)\Big) \xrightarrow[h\to\infty]{} +\infty.$$

But this is true: Indeed, the inequality

$$c(0) \geq c(p) + \nabla c(p) \cdot (-p)$$

and the superlinearity of c imply $\nabla c(p) \cdot (p/|p|) \to +\infty$ as $p \to \infty$, and then, with the notation $h_\varepsilon = h(1 - \varepsilon/|h|)$,

$$c(h) - c(h_\varepsilon) \geq \nabla c(h_\varepsilon) \cdot \frac{\varepsilon h}{|h|} = \varepsilon\, \nabla c(h_\varepsilon) \cdot \frac{h_\varepsilon}{|h_\varepsilon|} \Big(-\frac{\varepsilon}{|h|}\Big) \xrightarrow[|h|\to\infty]{} +\infty.$$

Example 10.21. If (M, g) is a Riemannian manifold with *nonnegative sectional curvature*, then (as recalled in the Third Appendix) there is a uniform upper bound $\nabla_x^2 (d(x_0, x)^2/2) \leq g_x$, and it follows that $c(x, y) = d(x, y)^2$ is semi-concave with a modulus $\omega(r) = r^2$. This condition of nonnegative curvature is quite restrictive, but there does not seem to be any good other geometric condition implying the semi-concavity of $d(x, y)^2$, uniformly in x and y.

I conclude this section with an open problem:

Open Problem 10.22. *Find simple sufficient conditions so that a cost deriving from a rather general Lagrangian on an unbounded Riemannian manifold will satisfy* $(\mathbf{H}\infty)$.

Differentiability of c-convex functions

Now we are back to optimal transport, and arrive at the core of the analysis of the Monge problem: the study of the regularity of c-convex functions. This includes c-subdifferentiability, subdifferentiability, and plain differentiability.

In all these theorems, M is a complete Riemannian manifold of dimension n, $\mathcal{X}$ is a closed subset of M such that the frontier $\partial\mathcal{X}$ (in M) is of dimension at most $n - 1$ (for instance it is locally a graph), and $\mathcal{Y}$ is an arbitrary Polish space. The cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is assumed to be continuous. The statements will be expressed in terms of the assumptions appearing in the previous section; these assumptions will be made for *interior* points, that is points which lie in the interior of $\mathcal{X}$ (viewed as a subset of M).

Theorem 10.23 (c-subdifferentiability of c-convex functions). *Assume that* $(\mathbf{H}\infty)$ *is satisfied. Let $\psi : M \to \mathbb{R} \cup \{+\infty\}$ be a c-convex function, and let Ω be the interior (in M) of its domain $\psi^{-1}(\mathbb{R})$. Then, $\psi^{-1}(\mathbb{R}) \setminus \Omega$ is a set of dimension at most $n-1$. Moreover, ψ is locally bounded and c-subdifferentiable everywhere in Ω. Finally, if K is a compact subset lying in Ω, then $\partial_c\psi(K)$ is itself compact.*

Theorem 10.24 (subdifferentiability of c-convex functions). *Assume that* **(Super)** *is satisfied. Let ψ be a c-convex function, and let x be an interior point of $\mathcal{X}$ (in M) such that $\partial_c\psi(x) \neq \emptyset$. Then ψ is subdifferentiable at x. In short:*

$$\partial_c\psi(x) \neq \emptyset \implies \nabla^-\psi(x) \neq \emptyset.$$

More precisely, for any $y \in \partial_c\psi(x)$, one has $-\nabla_x^+ c(x,y) \subset \nabla^-\psi(x)$.

Theorem 10.25 (differentiability of c-convex functions). *Assume that* **(Super)** *and* **(Twist)** *are satisfied, and let ψ be a c-convex function. Then*

(i) If **(Lip)** *is satisfied, then ψ is locally Lipschitz and differentiable in $\mathcal{X}$, apart from a set of zero volume; The same is true if* **(locLip)** *and* **(H∞)** *are satisfied.*

(ii) If **(SC)** *is satisfied, then ψ is locally semi-convex and differentiable in the interior (in M) of its domain, apart from a set of dimension at most $n-1$; and the boundary of its domain is also of dimension at most $n-1$. The same is true if* **(locSC)** *and* **(H∞)** *are satisfied.*

Proof of Theorem 10.23. Let $S = \psi^{-1}(\mathbb{R}) \setminus \partial\mathcal{X}$. (Here $\partial\mathcal{X}$ is the boundary of $\mathcal{X}$ in M, which by assumption is of dimension at most $n-1$.) We shall show that if $x \in S$ is such that T_xS is not included in a half-space, then ψ is bounded on a small ball around x. It will follow that x is in fact in the interior of Ω. So for each $x \in S \setminus \Omega$, T_xS will be included in a half-space, and it will follow that $S \setminus \Omega$ is of dimension at most $n-1$. Moreover, we shall have shown that ψ is locally bounded in Ω.

So let x be such that $\psi(x) < +\infty$, and T_xS is not included in a half-space. By assumption, there are points $z_1, \ldots, z_k$ in S, a small ball B around x, and a compact set $K \subset \mathcal{Y}$ such that for any $y \in \mathcal{Y} \setminus K$,

$$\inf_{w\in B} c(w,y) \geq \inf_{1\leq j\leq k} c(z_j,y).$$

Let ϕ be the c-transform of ψ. Then, for any $y \in \mathcal{Y} \setminus K$,

$$\phi(y) - \inf_{w\in B} c(w,y) \leq \phi(y) - \inf_{1\leq j\leq k} c(z_j,y) \leq \sup_{1\leq j\leq k} \psi(z_j).$$

So

$$\forall w \in B, \quad \forall y \in \mathcal{Y} \setminus K, \qquad \phi(y) - c(w,y) \leq \sup_{1\leq j\leq k} \psi(z_j).$$

When $y \in K$, we use the trivial bound $\phi(y) - c(w,y) \leq \psi(x) + c(x,y) - c(w,y)$. So all in all,

$$\forall w \in B, \quad \psi(w) = \sup_{y\in\mathcal{Y}} [\phi(y) - c(w,y)] \leq \max\left(\sup_{1\leq j\leq k} \psi(z_j), \ \sup_{y\in K} c(x,y) + \psi(x) - c(w,y) \right).$$

This shows that ψ is indeed bounded above on B. On the other hand, it is lower semi-continuous with values in $\mathbb{R} \cup \{+\infty\}$, and therefore also bounded below on a neighborhood of x. All in all, ψ is bounded in a neighborhood of x.

Next, let $x \in \Omega$, the goal is to show that $\partial_c\psi(x) \neq \emptyset$. Let U be a small neighborhood of x, on which $|\psi|$ is bounded by M. By assumption there is a compact set K, and a small ball B' in U, such that for all y outside K,

$$\forall z \in B', \quad c(z,y) - c(x,y) \leq -(2M+1).$$

Then, if y is outside of K, there is z such that $\psi(z) + c(z,y) \le c(x,y) - (M+1) \le \psi(x) + c(x,y) - 1$, and

$$\phi(y) - c(x,y) \le \inf_{z \in B'} \big[\psi(z) + c(z,y) - c(x,y)\big] \le \psi(x) - 1 = \sup_{y' \in \mathcal{Y}} [\phi(y') - c(x,y')] - 1.$$

Then the supremum of $\phi(y) - c(x,y)$ over all $\mathcal{Y}$ is the same as the supremum over only K. But this is a maximization problem for an upper semi-continuous function on a compact set, so it admits a solution. This solution is an element of $\partial_c \psi(x)$.

The same reasoning can be made with x replaced by w in a small neighborhood B of x, and then the conclusion is that $\partial_c \psi(w)$ is nonempty and contained in the compact set K, uniformly for $z \in B$. Then if $K' \subset \Omega$ is a compact set, we can cover it by a finite number of small open balls B_j such that $\partial_c \psi(B_j)$ is contained in a compact set K_j, so that $\partial_c \psi(K') \subset \cup K_j$. Since on the other hand $\partial_c \psi(K')$ is closed by the continuity of c, it follows that $\partial_c \psi(K')$ is compact. This concludes the proof of Theorem 10.23. □

Proof of Theorem 10.24. Let x be a point of c-subdifferentiability of ψ, and let $y \in \partial_c \psi(x)$. Let further

$$\phi(y) := \inf \big[\psi(x) + c(x,y)\big]$$

be the c-transform of ψ. By definition of c-subdifferentiability,

$$\psi(x) = \phi(y) - c(x,y). \tag{10.15}$$

Let x_ε be obtained from x by a small variation in the direction $w \in T_x M$, say $x_\varepsilon = \exp_x(\varepsilon w)$. From the definition of ϕ, one has of course

$$\psi(x_\varepsilon) \ge \phi(y) - c(x_\varepsilon, y). \tag{10.16}$$

Let further $p \in \nabla_x^+ c(x,y)$. By (10.15), (10.16) and the superdifferentiability of c, one obtains

$$\begin{aligned} \psi(x_\varepsilon) &\ge \phi(y) - c(x_\varepsilon, y) \\ &\ge \phi(y) - c(x,y) + \varepsilon \langle p, w \rangle + o(\varepsilon) \\ &= \psi(x) + \varepsilon \langle p, w \rangle + o(\varepsilon). \end{aligned}$$

where v denotes, as before, the velocity of a minimizing curve joining x to y. This shows that ψ is indeed subdifferentiable at x, with p as a subgradient. □

Proof of Theorem 10.25. (i) If c is locally Lipschitz as a function of x, with a uniform (in y) constant, then also $\psi(x) = \sup_y [\phi(y) - c(x,y)]$ is locally Lipschitz in x. Then it follows by Theorem 10.8(ii) that it is differentiable everywhere on the interior of $\mathcal{X}$, apart from a set of zero volume.

If c is only locally Lipschitz in x and y, but condition **(H∞)** is ensured, then for each compact set K in $\mathcal{X}$ there is a compact set K' in $\mathcal{Y}$ such that

$$\forall x \in K, \qquad \psi(x) = \sup_{y \in \partial_c \psi(x)} [\phi(y) - c(x,y)] = \sup_{y \in K'} [\phi(y) - c(x,y)].$$

The functions inside the supremum are uniformly Lipschitz when x stays in K and y stays in K', so the result of the supremum is again a locally Lipschitz function.

(ii) Let K be a compact subset of M, and let γ be a constant-speed geodesic whose image is included in K, then by assumption, for all y,

$$c(\gamma_t, y) \geq (1-t)\, c(\gamma_0, y) + t\, c(\gamma_1, y) - t(1-t)\, \omega\big(d(\gamma_0, \gamma_1)\big),$$

so

$$\begin{aligned}
\psi(\gamma_t) &= \sup_y \big[\phi(y) - c(\gamma_t, y)\big] \\
&\leq \sup_y \Big[\phi(y) - (1-t)\, c(\gamma_0, y) - t\, c(\gamma_1, y)\Big] + t(1-t)\omega\big(d(\gamma_0, \gamma_1)\big) \\
&= \sup_y \Big[(1-t)\, \phi(y) - (1-t)\, c(\gamma_0, y) + t\, \phi(y) - t\, c(\gamma_1, y)\Big] + t(1-t)\, \omega\big(d(\gamma_0, \gamma_1)\big) \\
&\leq (1-t)\, \sup_y \big[\phi(y) - c(\gamma_0, y)\big] + t\, \sup_y \big[\phi(y) - c(\gamma_1, y)\big] + t(1-t)\, \omega\big(d(\gamma_0, \gamma_1)\big) \\
&= (1-t)\, \psi(\gamma_0) + t\, \psi(\gamma_1) + t(1-t)\, \omega\big(d(\gamma_0, \gamma_1)\big).
\end{aligned}$$

So ψ inherits the semi-concavity modulus of c as semi-convexity modulus. Then, the conclusion follows from Proposition 10.12 and Theorem 10.8(iii). The localization argument is the same as in the proof of Statement (i). □

Applications to the Monge problem

The next theorem shows how to incorporate the previous information into the optimal transport.

Theorem 10.26 (Solution of the Monge problem II). *Let M be a Riemannian manifold, $\mathcal{X}$ a closed subset of M, with $\dim(\partial\mathcal{X}) \leq n-1$, and $\mathcal{Y}$ an arbitrary Polish space. Let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a continuous cost function, bounded below, and $\mu \in P(M)$, $\nu \in P(\mathcal{Y})$, such that the optimal cost $C(\mu, \nu)$ is finite. Assume that*

(i) c is superdifferentiable everywhere (Assumption **(Super)***);*

(ii) $\nabla_x c(x, \cdot)$ is injective on its domain of definition (Assumption **(Twist)***);*

(iii) Any c-convex function is differentiable μ-almost surely on its domain of c-subdifferentiability.

Then there exists a unique (in law) optimal coupling (x, y) of (μ, ν); it is deterministic, and there is a c-convex function ψ such that

$$\nabla\psi(x) + \nabla_x c(x, y) = 0 \qquad \text{almost surely.} \tag{10.17}$$

If moreover **(H∞)** *is satisfied, then*

(a) the equation (10.17) *characterizes the optimal coupling;*

(b) let Z be the set of points where ψ is differentiable; then one can define a continuous map $x \to T(x)$ on Z by the equation $T(x) \in \partial_c\psi(x)$, and

$$\mathrm{Spt}\, \nu = \overline{T(\mathrm{Spt}\, \mu)}. \tag{10.18}$$

Remark 10.27. As a corollary of this theorem, $\nabla\psi$ is uniquely determined μ-almost surely, since the random variable $\nabla\psi(x)$ has to coincide (in law) with $-\nabla_x c(x, y)$.

Remark 10.28. Assumption (iii) can be realized in a number of ways, depending on which part of Theorem 10.25 one wishes to use: For instance, it is true if c is Lipschitz on $\mathcal{X} \times \mathcal{Y}$ and μ is absolutely continuous; or if c is locally Lipschitz and μ, ν are compactly supported and μ is absolutely continuous; or if c is locally semi-concave and satisfies **(H∞)** and μ does not charge sets of dimension $n-1$; etc.

Example 10.29. All the assumptions of Theorem 10.26 are satisfied if $\mathcal{X} = M = \mathcal{Y}$ is compact and the Lagrangian L is C^2 and satisfies the classical assumptions of Definition 7.6.

Example 10.30. All the assumptions of Theorem 10.26 are satisfied if $\mathcal{X} = M = \mathcal{Y} = \mathbb{R}^n$, c is a C^1 strictly convex function with a bounded Hessian and μ does not charge sets of dimension $n-1$. Indeed, $\nabla_x c$ will be injective by strict convexity of c; and c will be uniformly semi-concave with a modulus Cr^2, so Theorem 10.25 applies to guarantee that c-convex functions are differentiable everywhere apart from a set of dimension at most $n-1$.

Example 10.31. All the assumptions of Theorem 10.26 are satisfied if $\mathcal{X} = M = \mathcal{Y}$ and $c(x,y) = d(x,y)^2$, and M has nonnegative sectional curvature; recall indeed Example 10.21. The same is true if M is compact.

Proof of Theorem 10.26. Let π be an optimal transference plan. From Theorem 5.9, there exists a pair of c-conjugate functions (ψ, ϕ) such that $\phi(y) - \psi(x) \le c(x,y)$ everywhere, with equality π-almost surely. Write again (10.2), at a point x of differentiability of ψ (x should be interior to $\mathcal{X}$, viewed as a subset of M), and choose $\widetilde{x} = \widetilde{x}(\varepsilon) = \gamma(\varepsilon)$, where $\dot\gamma(0) = w$; divide by $\varepsilon > 0$ and pass to the lim inf:

$$-\nabla\psi(x) \cdot w \le \liminf_{\varepsilon \to 0} \ \frac{c(\widetilde{x}(\varepsilon), y) - c(x,y)}{\varepsilon}. \tag{10.19}$$

It follows that $-\nabla\psi(x)$ is a subgradient of $c(\cdot, y)$ at x. But by assumption, there exists at least one supergradient of $c(\cdot, y)$ at x, say G. From Proposition 10.7, $c(\cdot, y)$ really is differentiable at x, with gradient $-\nabla\psi(x)$.

So (10.17) holds true, and then one can use assumption (iii) to derive $y = (\nabla_x c)^{-1}(x, -\nabla\psi(x))$, where $(\nabla_x c)^{-1}$ is the inverse of $x \longmapsto \nabla_x c(x,y)$, viewed as a function of y and defined on the set of x for which $\nabla_x c(x,y)$ exists.

So π is concentrated on the graph of $T : x \to (\nabla_x c)^{-1}(x, -\nabla\psi(x))$; or equivalently, $\pi = (\mathrm{Id}\,, T)_\# \mu$. Since this conclusion does not depend on the choice of π, but only on the choice of ψ, there is uniqueness of the optimal coupling π.

It remains to prove the last part of Theorem 10.26. From now on I shall assume that **(H∞)** is satisfied. Let π be a transference plan between μ and ν, and let ψ be a c-convex function such that (10.17) holds true.

Let Z be the set of differentiability points of ψ, and let $x \in Z$; in particular, x should be interior to $\mathcal{X}$ (in M), and should belong to the interior of the domain of ψ. By Theorem 10.23, there is some $y \in \partial_c\psi(x)$. Let G be a supergradient of $c(\cdot, y)$ at x; by Theorem 10.24, $-G \subset \{\nabla^-\psi(x)\} = \{\nabla\psi(x)\}$. It follows that $-\nabla\psi(x)$ is the only supergradient of $c(\cdot, y)$ at x (as in the beginning of the proof of the present theorem); so $c(\cdot, y)$ really is differentiable at x and $\nabla_x c(x,y) + \nabla\psi(x) = 0$. By injectivity of $\nabla_x c(x, \cdot)$, this equation determines $y = T(x)$ as a function of $x \in Z$. This proves the first part of (b), and also shows that $\partial_c\psi(x) = \{T(x)\}$ for any $x \in Z$.

Now, since π is concentrated on $Z \times \mathcal{Y}$, it follows from the equation (10.17) shows that π really is concentrated on the graph of T. A fortiori $\pi[\partial_c\psi] = 1$, so π is c-cyclically monotone, and therefore optimal by Theorem 5.9. This proves (a).

Next, let us prove that T is continuous on Z. Let $(x_k)_{k\in\mathbb{N}}$ be a sequence in Z, converging to $x \in Z$, and let $y_k = T(x_k)$. Assumption **(H∞)** and Theorem 10.23 imply that $\partial_c\psi$ transforms compact sets into compact sets; so the sequence $(y_k)_{k\in\mathbb{N}}$ takes values in a compact set, and up to extraction of a subsequence it converges to some $y' \in \mathcal{Y}$. By passing

to the limit in the inequality defining the c-subdifferential, we recover $y' \in \partial_c\psi(x)$. Since $x \in Z$, this determines $y' = T(x)$ uniquely, so the whole sequence $(T(x_k))_{k\in\mathbb{N}}$ converges to $T(x)$, and T is indeed continuous.

Equation (10.18) follows from the continuity of T. Indeed, the inclusion $\mathrm{Spt}\,\mu \subset T^{-1}(T(\mathrm{Spt}\,\mu))$ implies

$$\nu[T(\mathrm{Spt}\,\mu)] = \mu[T^{-1}(T(\mathrm{Spt}\,\mu))] \geq \mu[\mathrm{Spt}\,\mu] = 1;$$

so by definition of support, $\mathrm{Spt}\,\nu \subset \overline{T(\mathrm{Spt}\,\mu)}$. On the other hand, if $x \in \mathrm{Spt}\,\mu \cap Z$, let $y = T(x)$, and let $\varepsilon > 0$; by continuity of T there is $\delta > 0$ such that $T(B_\delta(x)) \subset B_\varepsilon(y)$, and then

$$\nu[B_\varepsilon(y)] = \mu\big[T^{-1}(B_\varepsilon(y))\big] \geq \mu\big[T^{-1}\big(T(B_\delta(x))\big)\big] \geq \mu[B_\delta(x)] > 0;$$

so $y \in \mathrm{Spt}\,\nu$. This shows that $T(\mathrm{Spt}\,\mu) \subset \mathrm{Spt}\,\nu$, and therefore $\overline{T(\mathrm{Spt}\,\mu)} \subset \mathrm{Spt}\,\nu$, as desired. This concludes the proof of (b). □

Removing the conditions at infinity

In this last section we shall see how to extend Theorem 10.26 to cover situations in which no control at infinity is assumed. The short answer is that it is sufficient to replace the *gradient* in (10.17) by an *approximate gradient*. (Actually a little bit more will be lost, see Remarks 10.33 and 10.34 below.)

Theorem 10.32 (Solution of the Monge problem without conditions at infinity). *Let M be a Riemannian manifold and $\mathcal{Y}$ an arbitrary Polish space. Let $c : M \times \mathcal{Y} \to \mathbb{R}$ be a continuous cost function, bounded below, and let $\mu \in P(M)$, $\nu \in P(\mathcal{Y})$, such that the optimal cost $C(\mu,\nu)$ is finite. Assume that*

(i) c is superdifferentiable everywhere (Assumption **(Super)***);*

(ii) $\nabla_x c(x,\cdot)$ is injective on its domain of definition (Assumption **(Twist)***);*

(iii) For any closed ball $B = B_{r]}(x_0)$ and any compact set $K \subset \mathcal{Y}$, the function c' defined on $B \times K$ by restriction of c is such that any c'-convex function on $B \times K$ is differentiable μ-almost surely;

(iv) μ is absolutely continuous with respect to the volume measure;

Then there exists a unique (in law) optimal coupling (x,y) of (μ,ν); it is deterministic, and satisfies the equation

$$\widetilde{\nabla}\psi(x) + \nabla_x c(x,y) = 0 \qquad \textit{almost surely.} \tag{10.20}$$

Remark 10.33. I don't know if (10.20) is a *characterization* of the optimal transport.

Remark 10.34. If Assumption (iv) is weakened into *(iv') μ gives zero mass to sets of dimension at most $n-1$*, then there is still uniqueness of the optimal coupling, and there is a c-convex ψ such that $y \in \partial_c\psi(x)$ almost surely; but it is not clear that equation (10.20) still holds. This uniqueness result is a bit more tricky than the previous one, and I shall postponed its proof to the next section (see Theorem 10.36).

Proof of Theorem 10.32. Let ψ be a c-convex function as in Theorem 5.9. Let π be an optimal transport π; according to Theorem 5.9, $\pi[\partial_c\psi] = 1$.

Let x_0 be any point in M, and y_0 any point in $\mathcal{Y}$. For any $\ell \in \mathbb{N}$, let B_ℓ be the closed ball $B_{\ell]}(x_0)$. Let also $(K_\ell)_{\ell\in\mathbb{N}}$ be an increasing sequence of compact sets in $\mathcal{Y}$, such that $\nu[\cup K_\ell] = 1$. The sets $B_\ell \times K_\ell$ fill up the whole of $M \times \mathcal{Y}$, up to a π-negligible set. Let c_ℓ be the restriction of c to $B_\ell \times K_\ell$.

If ℓ is large enough, then $\pi[B_\ell \times K_\ell] > 0$, so we can define

$$\pi_\ell := \frac{1_{B_\ell \times K_\ell}\, \pi}{\pi[B_\ell \times K_\ell]},$$

and then introduce the marginals μ_ℓ and ν_ℓ of π_ℓ. From Theorem 4.5, π_ℓ is optimal in the transport problem from μ_ℓ to ν_ℓ, with cost c_ℓ. By Theorem 5.16 we can find a c_ℓ-convex function ψ_ℓ which coincides with ψ μ_ℓ-almost surely, and actually on the whole of $S_\ell := \mathrm{proj}_M((\partial_c\psi) \cap (B_\ell \times K_\ell))$. (Note that S_ℓ is compact since $\partial_c\psi$ is closed.)

The union of all sets S_ℓ covers $\mathrm{proj}_M(\partial_c\psi)$, and therefore also $\mathrm{proj}_M(Spt(\pi))$, apart from a μ-negligible set. Let $\widetilde{S}_\ell$ be the set of points in S_ℓ at which S_ℓ has density 1; we know that $\widetilde{S}_\ell$ coincides with S_ℓ apart from a set of zero volume. So the union of all sets $\widetilde{S}_\ell$ still covers M, apart from a μ-negligible set.

By assumption (iii), ψ_ℓ is differentiable apart from a μ_ℓ-negligible set Z_ℓ. Moreover, by Theorem 10.26, the equation

$$\nabla\psi_\ell(x) + \nabla_x c(x, y) = 0 \tag{10.21}$$

determines the unique optimal coupling between μ_ℓ and ν_ℓ, for the cost c_ℓ. (Note that $\nabla_x c_\ell$ coincides with $\nabla_x c$ when x is in the interior of B_ℓ, and $\mu_\ell[\partial B_\ell] = 0$, so equation (10.21) does hold true μ_ℓ-almost surely.)

Now we can define our Monge coupling. Each set Z_ℓ is μ_ℓ-negligible, so it is also μ-negligible, and the union of all sets $\widetilde{S}_\ell \setminus Z_\ell$ still covers M, apart from a μ-negligible set. For each (x, y) in the support of π_ℓ, such that $x \in \widetilde{S}_\ell \setminus Z_\ell$, equation (10.21) holds true.

By definition of $\widetilde{S}$, ψ_ℓ coincides with ψ on a set which has density 1 at x, so $\nabla\psi_\ell(x) = \widetilde{\nabla}\psi(x)$, and then (10.21) becomes

$$\nabla_x c(x, y) + \widetilde{\nabla}\psi(x) = 0. \tag{10.22}$$

This equation is independent of ℓ, and for almost any (x, y) in the support of π there is ℓ large enough that (x, y) lies in the support of π_ℓ and $x \in \widetilde{S}_\ell \setminus Z_\ell$. The conclusion is that (10.22) holds true π-almost surely.

Since $\nabla_x c$ is injective, this equation determines $y = T(x)$ uniquely. The conclusion is that π is concentrated on the graph of the measurable map T defined by the equation

$$\nabla_x c(x, T(x)) + \widetilde{\nabla}\psi(x) = 0.$$

The uniqueness of the optimal coupling follows obviously. □

As an illustration of the use of Theorems 10.26 and 10.32, let us see how we can solve the Monge problem for the square distance on a Riemannian manifold. In the following statement, I say that M is *asymptotically flat* if all sectional curvatures σ_x at point x satisfy

$$\sigma_x \geq -\frac{C}{d(x_0, x)^2} \tag{10.23}$$

for some positive constant C and some $x_0 \in M$.

Theorem 10.35 (Monge problem for the square distance). *Let M be a smooth Riemannian manifold, and let $c(x,y) = d(x,y)^2$. Let μ, ν be two probability measures on M, such that the optimal cost between μ and ν is finite. If μ is absolutely continuous, then there is a unique solution of the Monge problem between μ and ν, and it can be written as*

$$y = T(x) = \exp_x\big(\widetilde{\nabla}\psi(x)\big), \tag{10.24}$$

where ψ is some $d^2/2$-convex function. The approximate gradient can be replaced by a true gradient if any one of the following conditions is satisfied:

(a) μ and ν are compactly supported;

(b) M has nonnegative sectional curvature;

(c) ν is compactly supported and M is asymptotically flat.

Proof. The general theorem is just a particular case of Theorem 10.32.

In case (a), we can apply Theorem 10.26(i) with $\mathcal{X} = B_{r]}(x_0) = \mathcal{Y}$, where r is large enough that $B_r(x_0)$ contains all geodesics that go from $\mathrm{Spt}\,\mu$ to $\mathrm{Spt}\,\nu$. Then the conclusion holds with some c'-convex function ψ, where c' is the restriction of c to $\mathcal{X} \times \mathcal{Y}$:

$$\psi(x) = \sup_{y \in B_{r]}(x_0)} \big[\phi(y) - \frac{d(x,y)^2}{2}\big].$$

To recover a true $d^2/2$-function, it suffices to set $\phi(y) = -\infty$ on $M \setminus \mathcal{Y}$, and let $\psi(x) = \sup_{y \in M}[\phi(y) - d(x,y)^2/2]$.

In case (b), all functions $d(\cdot, y)^2/2$ are uniformly semi-concave (as recalled in the Third Appendix), so Theorem 10.26(ii) applies.

In case (c), all functions $d(\cdot, y)^2/2$, where y varies in the support of ν, are uniformly semi-concave (as recalled in the Third Appendix), so We can choose $\mathcal{Y}$ to be a large closed ball containing the support of ν, and apply Theorem 10.26(ii) again. □

Removing the assumption of finite cost

In this section, I shall investigate situations where the total transport cost might be infinite. Unless the reader is specifically interested in such a situation, he or she is advised to skip this section which is quite tricky.

If $C(\mu, \nu) = +\infty$, there is no point in searching for an optimal transference plan. However it does make sense to look for c-cyclically monotone plans, that will be called **generalized optimal transference plans**.

Theorem 10.36 (Solution of the Monge problem without assumption of finite total cost). *Let $\mathcal{X}$ be a closed subset of a Riemannian manifold M such that $\dim(\partial\mathcal{X}) \leq n-1$, and let $\mathcal{Y}$ be an arbitrary Polish space. Let $c : M \times \mathcal{Y} \to \mathbb{R}$ be a continuous cost function, bounded below, and let $\mu \in P(M)$, $\nu \in P(\mathcal{Y})$. Assume that*

(i) c is superdifferentiable everywhere (Assumption **(Super)***);*

(ii) $\nabla_x c(x, \cdot)$ is injective on its domain of definition (Assumption **(Twist)***);*

(iii) c is locally semi-concave (Assumption **(locSC)***);*

(iv) μ does not give mass to sets of dimension at most $n-1$.

Then there exists a unique (in law) coupling (x,y) of (μ,ν) such that $\pi = \mathrm{law}\,(x,y)$ is c-cyclically monotone; moreover this coupling is deterministic. The measure π is called

the generalized optimal transference plan between μ and ν. Moreover, there is a c-convex function $\psi : M \to \mathbb{R} \cup \{+\infty\}$ such that $\pi[\partial_c \psi] = 1$.

• *If Assumption (iv) is reinforced into*

(iv') μ is absolutely continuous with respect to the volume measure,

then

$$\widetilde{\nabla}\psi(x) + \nabla_x c(x,y) = 0 \qquad \pi(dx\,dy)\text{-almost surely.} \tag{10.25}$$

• *If Assumption (iv) is left as it is, but one adds*

(v) the cost function satisfies **(H∞)** *or* **(SC)**,

then

$$\nabla\psi(x) + \nabla_x c(x,y) = 0 \qquad \pi(dx\,dy)\text{-almost surely,} \tag{10.26}$$

and this characterizes the generalized optimal transport. Moreover, one can define a continuous map $x \to T(x)$ on the set of differentiability points of ψ by the equation $T(x) \in \partial_c \psi(x)$, and then $\mathrm{Spt}\,\nu = \overline{T(\mathrm{Spt}\,\mu)}$.

Remark 10.37. Remark 10.33 applies also in this case.

Proof of Theorem 10.36. Let us first consider the existence problem. Let $(\mu_k)_{k\in\mathbb{N}}$ be a sequence of compactly supported probability measures converging weakly to μ; and similarly let $(\nu_k)_{k\in\mathbb{N}}$ be a sequence of compactly supported probability measures converging weakly to ν. For each index k, the total transport cost $C(\mu_k, \nu_k)$ is finite; let π_k be an optimal transference plan between μ_k and ν_k. By Theorem 5.9, π_k is c-cyclically monotone. By Lemma 4.3, the sequence $(\pi_k)_{k\in\mathbb{N}}$ converges, up to extraction, to some transference plan $\pi \in \Pi(\mu,\nu)$. By Theorem 5.18, π is c-cyclically monotone. By Step 3 of the proof of Theorem 5.9 (Rüschendorf's theorem), there is a c-convex ψ such that $\mathrm{Spt}(\pi) \subset \partial_c\psi$, in particular $\pi[\partial_c\psi] = 1$.

If μ is absolutely continuous, then we can proceed as in the proof of Theorem 10.32 to show that the coupling is deterministic and that (10.25) holds true π-almost surely.

In the case when **(H∞)** or **(SC)** is assumed, then we know that ψ is c-subdifferentiable everywhere in the interior of its domain; then we can proceed as in Theorem 10.26 to show that the coupling is deterministic, that (10.26) holds true, and that this equation implies $y \in \partial_c\psi(x)$; then if we prove the uniqueness of the generalized optimal transference plan this will show that (10.25) characterizes it.

So it all boils down to prove that *under Assumptions (i)–(iv), the generalized optimal transport is unique.* This will be much more technical, and the reader is advised to skip all the rest of the proof at first reading.

Let π be a generalized optimal coupling of μ and ν, and let ψ be a c-convex function such that $\mathrm{Spt}(\pi) \subset \partial_c\psi$. Let $z_0 \in \mathcal{X}$, let $B_\ell = B(z_0,\ell) \cup \mathcal{X}$; and let $(K_\ell)_{\ell\in\mathbb{N}}$ be an increasing sequences of compact subsets of $\mathcal{Y}$, such that $\nu[\cup K_\ell] = 1$. Let $Z_\ell := \pi_\ell[B_\ell \times K_\ell]$, $c_\ell := c|_{B_\ell\times K_\ell}$, $\pi_\ell := 1_{B_\ell\times K_\ell}\pi/Z_\ell$, $S_\ell := \mathrm{proj}_M(\mathrm{Spt}\,\pi_\ell)$; let also μ_ℓ and ν_ℓ be the two marginals of π_ℓ. It is easy to see that S_ℓ is still an increasing family of compact subsets of M, and that $\mu[\cup S_\ell] = 1$.

According to Theorem 5.16, there is a c_ℓ-convex function $\psi_\ell : B_\ell \to \mathbb{R}\cup\{+\infty\}$ which coincides with ψ on S_ℓ. Since c is locally semi-concave, the cost c_ℓ is uniformly semi-concave, and ψ_ℓ is differentiable on S_ℓ apart from a set of dimension $n-1$.

By **Besicovich's density theorem**, the set S_ℓ has μ-density 1 at μ-almost all $x \in S_\ell$; that is

$$\frac{\mu[S_\ell \cap B_r(x)]}{\mu[B_r(x)]} \xrightarrow[r\to 0]{} 1.$$

(The proof of this uses the fact that we are working on a Riemannian manifold; see the bibliographical notes for more information.)

Moreover, the transport plan π_ℓ induced by π on S_ℓ conicides with the deterministic transport associated with the map

$$T_\ell : x \to (\nabla_x c_\ell)^{-1}(x, -\nabla\psi_\ell(x)).$$

Since π is the nondecreasing limit of the transports π_ℓ, it follows that π itself is deterministic, and associated with the transport map T that sends x to $T_\ell(x)$ if $x \in S_\ell$. (This map is well-defined μ-almost surely.)

Let then

$$C_\ell := \Big\{x \in S_\ell;\ x \text{ is interior to } \mathcal{X};\quad S_\ell \text{ has } \mu\text{-density 1 at } x;$$
$$\forall k \geq \ell,\ \psi_k \text{ is differentiable at } x\Big\};\quad \nabla_x c(x, T(x)) + \nabla\psi_\ell(x) = 0\Big\}.$$

(Note: There is no reason for $\nabla\psi_\ell(x)$ to be an approximate gradient of ψ at x, because ψ_ℓ is assumed to coincide with ψ only on a set of μ-density 1 at x, not on a set of vol-density 1 at x.....)

The sets C_ℓ form a nondecreasing family of bounded Borel sets. Moreover, C_ℓ has been obtained from S_ℓ by deletion of a set of zero volume, and therefore of zero μ-measure. In particular, $\mu[\cup C_\ell] = 1$.

Let now $\widetilde{\pi}$ be another generalized optimal transference plan, and let $\widetilde{\psi}$ be a c-convex function with $\mathrm{Spt}(\widetilde{\pi}) \subset \partial_c\widetilde{\psi}$. We repeat the same construction as above with $\widetilde{\pi}$ instead of π, and get sequences $(\widetilde{Z}_\ell)_{\ell\in\mathbb{N}}$, $(\widetilde{\pi}_\ell)_{\ell\in\mathbb{N}}$, $(\widetilde{c}_\ell)_{\ell\in\mathbb{N}}$, $(\widetilde{\psi}_\ell)_{\ell\in\mathbb{N}}$, $(\widetilde{C}_\ell)_{\ell\in\mathbb{N}}$, such that the $\widetilde{C}_\ell$ form a nondecreasing family of bounded Borel sets with $\mu[\cup\widetilde{C}_\ell] = 1$, $\widetilde{\psi}_\ell$ coincides with $\widetilde{\psi}$ on $\widetilde{C}_\ell$. Also we find that $\widetilde{\pi}$ is deterministic and determined by the transport map $\widetilde{T}$, where $\widetilde{T}$ coincides with $\widetilde{T}_\ell$ on S_ℓ.

Next, the sets $C_\ell \cap \widetilde{C}_\ell$ also form a nondecreasing family of Borel sets, and $\mu[\cup(C_\ell \cap \widetilde{C}_\ell)] = \mu[(\cup C_\ell) \cap (\cup\widetilde{C}_\ell)] = 1$ (here the nondecreasing property was used in the first equality). Also $C_\ell \cap \widetilde{C}_\ell$ has μ-density 1 at each of its points.

Assume that $T \neq \widetilde{T}$ on a set of positive μ-measure; then there is some $\ell \in \mathbb{N}$ such that $\{T \neq \widetilde{T}\} \cap (C_\ell \cap \widetilde{C}_\ell)$ has positive μ-measure. This implies that $\{T_\ell \neq \widetilde{T}_\ell\} \cap (C_\ell \cap \widetilde{C}_\ell)$ has positive μ-measure, and then this implies that

$$\mu\big[\{\nabla\widetilde{\psi}_\ell \neq \nabla\psi\} \cap (C_\ell \cap \widetilde{C}_\ell)\big] > 0.$$

In the sequel, I shall fix such an ℓ.

Let x be a μ-Besicovich point of $E_\ell := (C_\ell \cap \widetilde{C}_\ell) \cap \{\nabla\widetilde{\psi}_\ell \neq \nabla\psi_\ell\}$, i.e. a point at which E_ℓ has μ-density 1. (Such a point exists since E_ℓ has positive μ-measure.) By adding a suitable constant to ψ, we may assume that $\widetilde{\psi}(x) = \psi(x)$. Since ψ_ℓ and $\widetilde{\psi}_\ell$ are semi-convex, we can apply the **implicit function theorem** to deduce that there is a small neighborhood of x in which the set $\{\psi_\ell = \widetilde{\psi}_\ell\}$ has dimension $n-1$. (See Corollary 10.44 in the last Appendix of this chapter.) Then, for r small enough, Assumption (iv) implies

$$\mu\Big[\{\psi_\ell \neq \widetilde{\psi}_\ell\} \cap B_r(x)\Big] = 0.$$

So at least one of the sets $\{\psi_\ell < \widetilde{\psi}_\ell\} \cap B_r(x)$ and $\{\psi_\ell > \widetilde{\psi}_\ell\} \cap B_r(x)$ has μ-measure at least $\mu[B_r(x)]/2$. Without loss of generality, I shall assume that this is the set $\{\psi_\ell > \widetilde{\psi}_\ell\}$; so

$$\mu\big[\{\psi_\ell > \widetilde{\psi}_\ell\} \cap B_{r_k}(x)\big] \geq \frac{\mu[B_{r_k}(x)]}{2}. \tag{10.27}$$

Next, ψ_ℓ coincides with ψ on the set S_ℓ, which has μ-density 1 at x, and similarly $\widetilde{\psi}_\ell$ coincides with $\widetilde{\psi}$ on a set of μ-density 1 at x. It follows that

$$\mu\big[\{\psi > \widetilde{\psi} \cap \{\psi_\ell > \widetilde{\psi}_\ell\} \cap B_{r_k}(x)\big] \geq \mu[B_{r_k}(x)]\ \Big(\frac{1}{2} - o(1)\Big). \tag{10.28}$$

Then since x is a Besicovich point of $\{\nabla\psi_\ell \neq \nabla\widetilde{\psi}_\ell\} \cap C_\ell \cap \widetilde{C}_\ell$,

$$\begin{aligned}\mu\big[\{\psi > \widetilde{\psi} \cap \{\psi_\ell > \widetilde{\psi}_\ell\} \cap \{\nabla\psi_\ell \neq \nabla\widetilde{\psi}_\ell\} \cap (C_\ell \cap \widetilde{C}_\ell) \cap B_{r_k}(x)\big]& \\ \geq \mu\big[\{\psi > \widetilde{\psi} \cap \{\psi_\ell > \widetilde{\psi}_\ell\} \cap B_{r_k}(x)\big] - \mu[B_{r_k}(x) \setminus (C_\ell \cap \widetilde{C}_\ell)]& \\ \geq \mu[B_{r_k}(x)]\ \Big(\frac{1}{2} - o(1) - o(1)\Big).&\end{aligned}$$

As a conclusion,

$$\forall r > 0 \quad \mu\Big[\{\psi > \widetilde{\psi}\} \cap \{\psi_\ell > \widetilde{\psi}_\ell\} \cap \{\nabla\psi_\ell \neq \nabla\widetilde{\psi}_\ell\} \cap (C_\ell \cap \widetilde{C}_\ell) \cap B_{r_k}(x)\Big] > 0. \tag{10.29}$$

Let now

$$A := \{\psi > \widetilde{\psi}\}.$$

The proof will result from the next two claims:

Claim 1: $\widetilde{T}^{-1}(T(A)) \subset A$;

Claim 2: The set $\{\psi_\ell > \widetilde{\psi}_\ell\} \cap (C_\ell \cap \widetilde{C}_\ell) \cap \{\nabla\psi_\ell \neq \nabla\widetilde{\psi}_\ell\} \cap \widetilde{T}^{-1}(T(A))$ lies a positive distance away from x.

Let us postpone the proofs of these claims for a while, and see how the theorem follows from them. Let $S \subset A$ be defined by $S := \{\psi > \widetilde{\psi} \cap \{\psi_\ell > \widetilde{\psi}_\ell\} \cap \{\nabla\psi_\ell \neq \nabla\widetilde{\psi}_\ell\} \cap (C_\ell \cap \widetilde{C}_\ell)$, and let

$$r := d\big(x, S \cap \widetilde{T}^{-1}(T(A))\big)/2.$$

Obviously, $\mu[S \cap B(x,r) \cap \widetilde{T}^{-1}(T(A))] = 0$. On the other hand, r is positive by Claim 2, so $\mu[S \cap B(x,r)] > 0$ by (10.29). Then

$$\mu\big[A \setminus \widetilde{T}^{-1}(T(A))\big] \geq \mu\big[S \cap B(x,r) \setminus \widetilde{T}^{-1}(T(A))\big] = \mu[S \cap B(x,r)] > 0.$$

Since $\widetilde{T}^{-1}(T(A)) \subset A$ by Claim 1, this implies

$$\mu[\widetilde{T}^{-1}(T(A))] < \mu[A]. \tag{10.30}$$

But then, we can write

$$\mu[A] \leq \mu[T^{-1}(T(A))] = \nu[T(A)] = \nu[\widetilde{T}(A)] = \mu[\widetilde{T}^{-1}(T(A))],$$

which contradicts (10.30). So it all boils down to establishing Claims 1 and 2 above.

Proof of Claim 1: Let $x \in \widetilde{T}^{-1}(T(A))$. Then there exists $y \in A$ such that $T(y) = \widetilde{T}(x)$. Recall that $T(y) \in \partial_c\psi(y)$ and $\widetilde{T}(x) \in \partial_c\psi(x)$. By using the definition of the c-subdifferential and the assumptions, we can write the following chain of identities and inequalities:

$$\begin{aligned}\widetilde{\psi}(x) + c(x, \widetilde{T}(x)) &\le \widetilde{\psi}(y) + c(y, \widetilde{T}(x)) \\ &= \widetilde{\psi}(y) + c(y, T(y)) \\ &< \psi(y) + c(y, T(y)) \\ &\le \psi(x) + c(x, T(y)) \\ &= \psi(x) + c(x, \widetilde{T}(x)).\end{aligned}$$

It follows that $\widetilde{\psi}(x) < \psi(x)$, i.e. $x \in A$. This proves Claim 1.

Proof of Claim 2: Assume that this claim is false; then there is a sequence $(x_k)_{k\in\mathbb{N}}$, valued in $\{\psi_\ell > \widetilde{\psi}_\ell\} \cap (C_\ell \cap \widetilde{C}_\ell) \cap \widetilde{T}^{-1}(T(A))$, such that $x_k \to x$. For each k, there is $y_k \in A$ such that $\widetilde{T}(x_k) = T(y_k)$. On $C_\ell \cap \widetilde{C}_\ell$, the transport T coincides with T_ℓ and the transport $\widetilde{T}$ with $\widetilde{T}_\ell$, so $\widetilde{T}(x_k) \in \partial_c \psi_\ell(x_k)$ and $T(y_k) \in \partial_c \psi_\ell(y_k)$; then we can write, for any $z \in M$,

$$\begin{aligned}\psi_\ell(z) &\ge \psi_\ell(y_k) + c(y_k, T(y_k)) - c(z, T(y_k)) \\ &= \psi_\ell(y_k) + c(y_k, \widetilde{T}(x_k)) - c(z, \widetilde{T}(x_k)) \\ &> \widetilde{\psi}_\ell(y_k) + c(y_k, \widetilde{T}(x_k)) - c(z, \widetilde{T}(x_k)) \\ &\ge \widetilde{\psi}_\ell(x_k) + c(x_k, \widetilde{T}(x_k)) - c(z, \widetilde{T}(x_k))\end{aligned}$$

Since ψ_ℓ is differentiable at x and since c is locally semi-concave by assumption, we can expand the right-hand side and obtain

$$\begin{aligned}\psi_\ell(z) \ge \widetilde{\psi}_\ell(x_k) + c_\ell(x_k, \widetilde{T}(x_k)) - c_\ell(z, \widetilde{T}(x_k)) = \nabla\widetilde{\psi}_\ell(x) + \nabla\widetilde{\psi}_\ell(x)\cdot(x_k - x) + o(|x_k - x|) \\ - \nabla_x c(x_k, \widetilde{T}(x_k))\cdot(z - x_k) + o(|z - x_k|), \quad (10.31)\end{aligned}$$

where $o(|z-x_k|)$ in the last line is uniform in k. (Here I have cheated by pretending to work in $\mathbb{R}^n$ rather than on a Riemannian manifold, but all this is purely local, and invariant under diffeomorphism; so there is really no problem to make sense of these formulas when z is close enough to x_k.) Recall that $\nabla_x c(x_k, \widetilde{T}(x_k)) + \nabla\widetilde{\psi}_\ell(x_k) = 0$; so (10.31) can be rewritten as

$$\psi(z) \ge \widetilde{\psi}_\ell(x_k) + \nabla\widetilde{\psi}_\ell(x) + \nabla\widetilde{\psi}_\ell(x)\cdot(x_k - x) + o(|x_k - x|) + \nabla\widetilde{\psi}_\ell(x_k)\cdot(z - x_k) + o(|z - x_k|). \quad (10.32)$$

Then we can pass to the limit as $k \to \infty$, remembering that $\nabla\widetilde{\psi}_\ell$ is continuous (because ψ_ℓ is semi-convex), and get

$$\begin{aligned}\psi_\ell(z) &\ge \widetilde{\psi}_\ell(x) + \nabla\widetilde{\psi}_\ell(x)\cdot(z - x) + o(|z - x|) \quad (10.33)\\ &= \psi_\ell(x) + \nabla\widetilde{\psi}_\ell(x)\cdot(z - x) + o(|z - x|).\end{aligned}$$

On the other hand, since ψ_ℓ is differentiable at x, we have

$$\psi_\ell(z) = \psi_\ell(x) + \nabla\psi_\ell(x)\cdot(z - x) + o(|z - x|).$$

Combining this with (10.33), we see that

$$(\nabla\widetilde{\psi}_\ell(x) - \nabla\psi_\ell(x))\cdot(z - x) \le o(|z - x|),$$

which is possible only if $\nabla\widetilde{\psi}_\ell(x) - \nabla\psi_\ell(x) = 0$. But this contradicts the definition of x. So Claim 2 should hold true, and this concludes the proof of Theorem 10.36. □

The next Corollary of Theorem 10.36 is exactly similar to Theorem 10.35 except that the classical Monge problem (search of a transport of minimum cost) has been replaced by the generalized Monge problem (search of a c-monotone transport).

Corollary 10.38 (Generalized Monge problem for the square distance). *Let M be a smooth Riemannian manifold, and let $c(x,y) = d(x,y)^2$. Let μ, ν be two probability measures on M.*

- If μ gives zero mass to sets of dimension at most $n-1$, then there is a unique solution $x \to T(x)$ of the generalized Monge problem between μ and ν.

- If μ is absolutely continuous, then this solution can be written

$$y = T(x) = \exp_x\big(\widetilde{\nabla}\psi(x)\big), \tag{10.34}$$

where ψ is some $d^2/2$-convex function.

- If M has nonnegative sectional curvature, or ν is compactly supported and M satisfies (10.23)*, then equation* (10.34) *still holds, but in addition the approximate gradient in that equation can be replaced by a true gradient.*

First Appendix: A little bit of Geometric Measure Theory

The geometric counterpart of differentiability is of course the approximation of a set S by a tangent plane, or hyperplane, or more generally by a tangent d-dimensional space, if d is the dimension of S.

If S is smooth, then there is no ambiguity on its dimension (a curve has dimension 1, a surface has dimension 2, etc.) and the tangent space always exists. But if S is not smooth, this might not be the case, at least not in the usual sense. The notion of **tangent cone** (sometimes called contingent cone) often remedies this problem; it is naturally associated with the notion of **countable d-rectifiability**, which acts as a replacement for the notion of "dimension d". Here below I shall recall some of the basic results about these concepts.

Definition 10.39 (tangent cone). *If S is an arbitrary subset of $\mathbb{R}^n$, and $x \in \overline{S}$, then the tangent cone T_xS to S at x is defined as*

$$T_xS := \left\{ \lim_{k\to\infty} \frac{x_k - x}{t_k};\quad x_k \in S,\ x_k \to x,\ t_k > 0,\ t_k \to 0 \right\}.$$

The dimension of this cone is defined as the dimension of the vector space that it generates.

Definition 10.40 (countable rectifiability). *Let S be a subset of $\mathbb{R}^n$, and let $d \in [0,n]$ be an integer. Then S is said to be countably d-rectifiable if $S \subset \bigcup_{k\in\mathbb{N}} f_k(D_k)$, where each f_k is Lipschitz on a measurable subset D_k of $\mathbb{R}^d$. In particular, it has Hausdorff dimension at most d.*

The next theorem summarizes two results which were useful in the present chapter:

Theorem 10.41 (sufficient conditions for countable rectifiability).

(i) Let S be a measurable set in $\mathbb{R}^n$, such that T_xS has dimension at most d for all $x \in S$. Then S is countably d-rectifiable.

(ii) Let S be a measurable set in $\mathbb{R}^n$, such that T_xS is included in a half-space, for each $x \in \partial S$. Then ∂S is countably $(n-1)$-rectifiable.

Proof of Theorem 10.41. Let us start with Statement (i). For each $x \in S$, let π_x stand for the orthogonal projection on T_xS, and let $\pi_x^\perp = \mathrm{Id} - \pi_x$ stand for the orthogonal projection on $(T_xS)^\perp$. I claim that

$$\forall x \in S, \quad \exists r > 0; \quad \forall y \in S, \quad |x-y| \le r \Longrightarrow |\pi_x^\perp(x-y)| \le |\pi_x(x-y)|. \tag{10.35}$$

Indeed, assume that (10.35) is false. Then there is $x \in S$, and there is a sequence $(y_k)_{k\in\mathbb{N}}$ such that $|x-y_k| \le 1/k$ and yet $|\pi_x^\perp(x-y)| > |\pi_x(x-y)|$, or equivalently

$$\left|\pi_x^\perp\left(\frac{x-y_k}{|x-y_k|}\right)\right| > \left|\pi_x\left(\frac{x-y_k}{|x-y_k|}\right)\right|. \tag{10.36}$$

Up to extraction of a subsequence, we may assume that $w_k := (x-y_k)/|x-y_k|$ converges to $\theta \in T_xS$ with $|\theta| = 1$. Then $|\pi_x w_k| \to 1$ and $|\pi_x^\perp w_k| \to 0$, which is in contradiction with (10.36). So (10.35) is true.

Next, for each $k \in \mathbb{N}$, let

$$S_k := \Big\{x \in S; \quad \text{property (10.35) holds true for } |x-y| \le 1/k\Big\}.$$

It is clear that the sets S_k cover S, so it is sufficient to prove the d-rectifiability of S_k for a given k.

Let $\delta > 0$ be small enough ($\delta < 1/2$ will do). Let Π_d be the set of all orthogonal projections on d-dimensional linear spaces. Since Π_d is compact, we can find a finite family $(\pi_1, \dots, \pi_N)$ of such orthogonal projections, such that for any $\pi \in \Pi_d$ there is $j \in \{1, \dots, N\}$ with $\|\pi - \pi_j\| \le \delta$, where $\|\cdot\|$ stands for the operator norm. So the set S_k is covered by the sets

$$S_{k\ell} := \Big\{x \in S_k; \quad \|\pi_x - \pi_\ell\| \le \delta\Big\}.$$

To prove the theorem, it suffices to prove that $S_{k\ell}$ is locally rectifiable. We shall show that

$$x, x' \in S_{k\ell},\ |x-x'| \le \frac{1}{k} \Longrightarrow |\pi_\ell^\perp(x-x')| \le L\,|\pi_\ell(x-x')|, \qquad L = \frac{1+2\delta}{1-2\delta}; \tag{10.37}$$

it will follow that the intersection of $S_{k\ell}$ with a ball of diameter $1/k$ is contained in an L-Lipschitz graph over $\pi_\ell(\mathbb{R}^n)$; and then the conclusion will follow immediately.

To prove (10.37), note that, if π and π' are any two orthogonal projections, then (since $\pi^\perp = \mathrm{Id} - \pi$), $\|\pi^\perp - (\pi')^\perp\| = \|\pi - \pi'\|$. So

$$\begin{aligned}
|\pi_\ell^\perp(x-x')| &\le |(\pi_\ell^\perp - \pi_x^\perp)(x-x')| + |\pi_x^\perp(x-x')| \\
&\le |(\pi_\ell - \pi_x)(x-x')| + |\pi_x(x-x')| \\
&\le |(\pi_\ell - \pi_x)(x-x')| + |\pi_\ell(x-x')| + |(\pi_\ell - \pi_x)(x-x')| \\
&\le |\pi_\ell(x-x')| + 2\delta|x-x'| \\
&\le (1+2\delta)|\pi_\ell(x-x')| + 2\delta|\pi_\ell^\perp(x-x')|.
\end{aligned}$$

This establishes (10.37).

Now let us turn to Statement (ii). Let F be a finite set in S^{n-1} such that the balls $(B_{1/8}(\nu))_{\nu\in F}$ cover S^{n-1}. I claim that

$$\forall x \in \partial S, \quad \exists r > 0, \quad \exists \nu \in F, \quad \forall y \in \partial S \cap B_r(x), \quad \langle y-x, \nu\rangle \le \frac{|y-x|}{2}. \tag{10.38}$$

Indeed, otherwise there is $x \in \partial S$ such that for all $k \in \mathbb{N}$ and for all $\nu \in F$ there is $y_k \in \partial S$ such that $|y_k - x| \le 1/k$ and $\langle y_k - x, \nu\rangle > |y_k - x|/2$. By assumption there is $\xi \in S^{n-1}$ such that

$$\forall \zeta \in T_x S, \quad \langle \xi, \zeta\rangle \le 0.$$

Let $\nu \in F$ be such that $|\xi - \nu| < 1/8$ and let $(y_k)_{k\in\mathbb{N}}$ be a sequence as above. Since $y_k \in \partial S$ and $y_k \neq x$, there is $y'_k \in S$ such that $|y_k - y'_k| < |y_k - x|/8$. Then

$$\langle y'_k - x, \xi\rangle \ge \langle y_k - x, \nu\rangle - |y_k - x|\,|\xi - \nu| - |y - y'_k| \ge \frac{|y_k - x|}{4} \ge \frac{|x - y'_k|}{8}.$$

So

$$\left\langle \frac{y'_k - x}{|y'_k - x|}, \xi \right\rangle \ge \frac{1}{8}. \tag{10.39}$$

Up to extraction of a subsequence, $(y'_k - x)/|y'_k - x|$ converges to some $\zeta \in T_x S$, and then by passing to the limit in (10.39) we have $\langle \zeta, \xi\rangle \ge 1/8$. But by definition, ξ is such that $\langle \zeta, \xi\rangle \le 0$ for all $\zeta \in T_x S$. This contradiction establishes (10.38).

As a consequence, ∂S is included in the union of all sets $A_{1/k,\nu}$, where $k \in \mathbb{N}$, $\nu \in F$, and

$$A_{r,\nu} := \left\{ x \in \partial S; \quad \forall y \in \partial S \cap B_r(x), \quad \langle y - x, \nu\rangle \le \frac{|y - x|}{2} \right\}.$$

To conclude the proof of the theorem it is sufficient to show that each $A_{r,\nu}$ is locally the image of a Lipschitz function defined on a subset of an $(n-1)$-dimensional space.

So let $r > 0$ and $\nu \in F$ be given, let $x_0 \in A_{r,\nu}$, and let π be the orthogonal projection of $\mathbb{R}^n$ to $\nu^\perp$. (Explicitly, $\pi(x) = x - \langle x, \nu\rangle \nu$.) We shall show that on $D := A_{r,\nu} \cap B_{r/2}(x_0)$, π is injective and its inverse (on $\pi(D)$) is Lipschitz. To see this, first note that for any two $x, x' \in D$, one has $x' \in B_r(x)$, so, by definition of $A_{r,\nu}$, $\langle x' - x, \nu\rangle \le |x' - x|/2$. By symmetry, also $\langle x - x', \nu\rangle \le |x - x'|/2$, so in fact

$$\big|\langle x - x', \nu\rangle\big| \le \frac{|x - x'|}{2}.$$

Then if $z = \pi(x)$ and $z' = \pi(x')$,

$$|x - x'| \le |z - z'| + \big|\langle x, \nu\rangle - \langle x', \nu\rangle\big| \le |z - z'| + \frac{|x - x'|}{2},$$

so $|x - x'| \le 2\,|z - z'|$. This concludes the proof. □

Second Appendix: Nonsmooth Implicit Function Theorem

Let M be an n-dimensional smooth Riemannian manifold, and $x_0 \in M$. I shall say that a set $M' \subset M$ is a k-dimensional C^r graph (resp. k-dimensional Lipschitz graph) in a neighborhood of x_0 if there are

(i) a smooth system of coordinates around x_0, say

$$x = \zeta(x', y),$$

where ζ is a smooth diffeomorphism from an open subset U of $\mathbb{R}^k \times \mathbb{R}^{n-k}$, into a neighborhood O of x_0;

(ii) a C^r (resp. Lipschitz) function $\varphi : O' \to \mathbb{R}^{n-k}$, where O' is an open subset of $\mathbb{R}^k$;

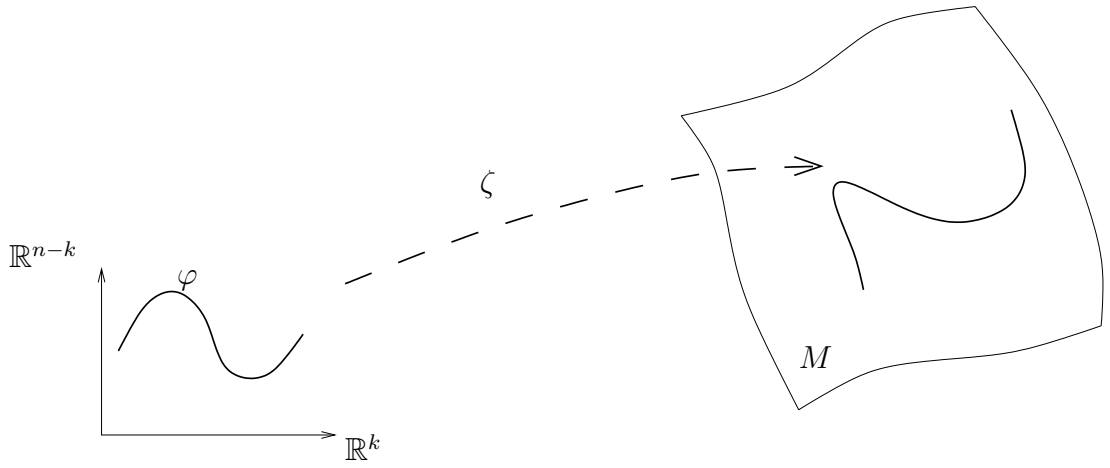


Fig. 10.2. k-dimensional graph

such that for all $x \in O$,

$$x \in M' \iff y = \varphi(x').$$

The following statement is a consequence of the classical implicit function theorem: *If $f : M \to \mathbb{R}$ is of class C^r $(r \geq 1)$, $f(x_0) = 0$ and $\nabla f(x_0) \neq 0$, then the set $\{f = 0\} = f^{-1}(0)$ is an $(n-1)$-dimensional C^r graph in a neighborhood of x_0.*

In this Appendix I shall consider a *nonsmooth* version of this theorem. The following notion will be useful.

Definition 10.42 (Clarke subdifferential). *Let f be a continuous real-valued function defined on an open subset U of a Riemannian manifold. For each $x \in U$, define $\partial f(x)$ as the convex hull of all limits of sequences $\nabla f(x_k)$, where all x_k are differentiability points of f and $x_k \to x$. In short:*

$$\partial f(x) = \overline{\mathrm{Conv}} \left\{ \lim_{x_k \to x} \nabla f(x_k) \right\}.$$

Further recall that if $(A_i)_{1 \leq i \leq m}$ are subsets of a vector space, then $\sum A_i = \{\sum_i a_i;\ a_i \in A_i\}$.

Theorem 10.43 (Nonsmooth Implicit Function Theorem). *Let $(f_i)_{1 \leq i \leq m}$ be real-valued Lipschitz functions defined in an open set U of an n-dimensional Riemannian manifold, and let $x_0 \in U$ be such that*

(a) $\sum f_i(x_0) = 0$;

(b) $0 \notin \sum \partial f_i(x_0)$.

Then $\{\sum f_i = 0\}$ is an $(n-1)$-dimensional Lipschitz graph around x_0.

Corollary 10.44 (Implicit function theorem for two semi-convex functions). *Let ψ and $\widetilde{\psi}$ be two semi-convex functions defined in an open set U of an n-dimensional Riemannian manifold M, and let $x_0 \in U$ be such that ψ, $\widetilde{\psi}$ are differentiable at x_0, and*

$$\psi(x_0) = \widetilde{\psi}(x_0); \qquad \nabla\psi(x_0) \neq \nabla\widetilde{\psi}(x_0).$$

Then there is a neighborhood V of x_0 such that $\{\psi = \widetilde{\psi}\} \cap V$ is an $(n-1)$-dimensional Lipschitz graph; in particular, it has Hausdorff dimension exactly $n-1$.

Proof of Corollary 10.44. Let $f_1 = \psi$, $f_2 = -\widetilde{\psi}$. Since f_1 is semi-convex and f_2 is semi-concave, both ∇f_1 and ∇f_2 are continuous on their respective domain of definition. So $\partial f_i(x_0) = \{\nabla f_i(x_0)\}$ $(i = 1, 2)$ and $\sum \partial f_i(x_0) = \{\sum \nabla f_i(x_0)\}$ does not contain 0. Then Theorem 10.43 implies the conclusion. □

Proof of Theorem 10.43. The statement is purely local and invariant under C^1 diffeomorphism, so we might assume that we are working in $\mathbb{R}^n$. For each i, $\partial f_i(x_0) \subset B(0, \|f_i\|_{\mathrm{Lip}}) \subset \mathbb{R}^n$, so $\partial f_i(x_0)$ is a compact convex subset of $\mathbb{R}^n$; then also $\sum \partial f_i(x_0)$ is compact and convex, and by assumption does not contain 0. By the Hahn–Banach theorem, there are $v \in \mathbb{R}^n$ and $\alpha > 0$ such that

$$\langle p, v\rangle \geq \alpha \qquad \text{for all } p \in \sum \partial f_i(x_0). \tag{10.40}$$

Then there is a neighborhood V of x_0 such that $\sum \langle \nabla f_i(x), v\rangle \geq \alpha/2$ at all points x where all functions f_i are differentiable. (Otherwise there would be a sequence $(x_k)_{k\in\mathbb{N}}$, converging to x_0, such that $\sum \langle \nabla f_i(x_k), v\rangle < \alpha/2$, but then up to extraction of a subsequence we would have $\nabla f_i(x_k) \to p_i \in \partial f_i(x_0)$, so $\langle \sum p_i, v\rangle \leq \alpha/2 < \alpha$, which would contradict (10.40).)

Without loss of generality, we may assume that $x_0 = 0$, $v = (e_1, 0, \ldots, 0)$, $V = (-\beta, \beta) \times B(0, r_0)$, where the latter ball is a subset of $\mathbb{R}^{n-1}$ and $r_0 \leq (\alpha\beta)/(4 \sum \|f_i\|_{\mathrm{Lip}})$. Let further

$$Z' := \Big\{ y \in B(0, r_0) \subset \mathbb{R}^{n-1};\ \lambda_1\big[\{t \in (-\beta, \beta);\ \exists i;\ \nabla f_i(t, y) \text{ does not exist}\}\big] > 0 \Big\};$$

$$Z = (-\beta, \beta) \times Z'; \qquad D = V \setminus Z.$$

I claim that $\lambda_n[Z] = 0$. To prove this it is sufficient to check that $\lambda_{n-1}[Z'] = 0$. But Z' is the noincreasing limit of $(Z'_\ell)_{\ell\in\mathbb{N}}$, where

$$Z'_\ell = \Big\{ y \in B(0, r_0);\ \lambda_1\big[\{t \in (-\beta, \beta);\ \exists i;\ \nabla f_i(t, y) \text{ does not exist}\}\big] \geq 1/\ell \Big\}.$$

By Fubini's theorem,

$$\lambda_n\Big[\big\{x \in O;\ \nabla f_i(x) \text{ does not exist for some } i\big\}\Big] \geq (\lambda_{n-1}[Z'_\ell]) \times (1/\ell);$$

and the left-hand side is equal to 0 since all f_i are differentiable almost everywhere. It follows that $\lambda_{n-1}[Z'_\ell] = 0$, and by taking the limit $\ell \to \infty$ we obtain $\lambda_{n-1}[Z'] = 0$.

Let $f = \sum f_i$, and let $\partial_1 f = \langle \nabla f, v\rangle$ stand for its partial derivative with respect to the first coordinate. The first step of the proof has shown that $\partial_1 f(x) \geq \alpha/2$ at each point x where all functions f_i are differentiable. So, for each $y \in O' \setminus Z'$, the function $t \to f(t, y)$ is Lipschitz and differentiable λ_1-almost everywhere on $(-\beta, \beta)$, and it satisfies $f'(t, y) \geq \alpha/2$. It follows that for all $t, t' \in (-\beta, \beta)$,

$$t < t' \Longrightarrow \quad f(t', y) - f(t, y) \geq (\alpha/2)\,(t' - t). \tag{10.41}$$

Since holds true for all $((t, y), (t', y))$ in $D \times D$. Since $Z = V \setminus D$ has zero Lebesgue measure, it follows that D is dense in V, so (10.41) extends to all $((t, y), (t', y)) \in V$.

For all $y \in B(0, r_0)$, inequality (10.41), combined with the estimate

$$|f(0, y)| = |f(0, y) - f(0, 0)| \leq \|f\|_{\mathrm{Lip}}\,|y| \leq \frac{\alpha\beta}{4},$$

guarantees that the equation $f(t, y) = 0$ has exactly one solution $t = \varphi(y)$ in $(-\beta, \beta)$.

It only remains to check that φ is Lipschitz on O'. Let $y, z \in O'$, then $f(\varphi(y), y) = f(\varphi(z), z) = 0$, so

$$\big|f(\varphi(y),y) - f(\varphi(z),y)\big| = \big|f(\varphi(z),z) - f(\varphi(z),y)\big|. \tag{10.42}$$

Since the first partial derivative of f is no less than $\alpha/2$, the left-hand side of (10.42) is bounded below by $(\alpha/2)|\varphi(y) - \varphi(z)|$, while the right-hand side is bounded above by $\|f\|_{\mathrm{Lip}}\,|z-y|$. The conclusion is that

$$|\varphi(y) - \varphi(z)| \le \frac{2\,\|f\|_{\mathrm{Lip}}}{\alpha}\,|z-y|,$$

so φ is indeed Lipschitz. □

Third Appendix: Curvature and the Hessian of the squared distance

The practical verification of the uniform semi-concavity of a given cost function $c(x,y)$ might be a very complicated task in general. In the particular case when $c(x,y) = d(x,y)^2$, then this problem can be related to the **sectional curvature** of the Riemannian manifold. In this Appendix I shall recall some results about these links, some of them well-known, other ones more confidential. The reader who does not know about sectional curvature can skip this Appendix, or take a look at Chapter 14 first. In any case the proofs in this Appendix will use some basic results in Riemannian geometry.

If $M = \mathbb{R}^n$ is the Euclidean space, then $d(x,y) = |x-y|$ and there is the simple formula

$$\nabla_x^2 \left(\frac{|x-y|^2}{2} \right) = I_n,$$

where the right-hand side is just the identity operator on $T_x\mathbb{R}^n = \mathbb{T}^n$.

If M is an arbitrary Riemannian manifold, there is no simple formula for $\nabla_x^2 d(x,y)^2/2$, and this operator will in general not be defined in the sense that it can take eigenvalues $-\infty$ if x and y are conjugate point. However, there is still a recipe to estimate $\nabla_x^2 d(x,y)^2/2$ *from above*, and thus derive estimates of semi-concavity for $d^2/2$.

So let x and y be any two points in M, and let γ be a minimizing geodesic joining y to x, parametrized by arc length; so $\gamma(0) = y$, $\gamma(d(x,y)) = x$. Let $H(t)$ stand for the Hessian operator of $x \to d(x,y)^2/2$ at $x = \gamma(t)$. On $[0, d(x,y))$ the operator $H(t)$ is well-defined (since the geodesic is minimizing, it is only at $t = d(x,y)$ that eigenvalues $-\infty$ may appear). It starts at $H(0) = \mathrm{Id}$ and then its eigenvectors and eigenvalues vary smoothly at t varies in $(0, d(x,y))$.

The unit vector $\dot\gamma(t)$ is an eigenvector of $H(t)$, associated with the eigenvalue $+1$. The problem is to bound the eigenvalues in the orthogonal subspace $S(t) = (\dot\gamma)^\perp \subset T_{\gamma(t)}M$. So let $(e_2, \ldots, e_n)$ be an orthonormal basis of $S(0)$, and let $(e_2(t), \ldots, e_n(t))$ be obtained by parallel transport of $(e_2, \ldots, e_n)$ along γ; it remains an orthonormal basis of $S(t)$. To achieve our goal, it is sufficient to bound above the quantities $h(t) = \langle H(t)\cdot e_i(t), e_i(t)\rangle_{\gamma(t)}$, where i is arbitrary in $\{2, \ldots, n\}$.

Since $H(0)$ is the identity, we have $h(0) = 1$. To get a differential equation on $h(t)$, we can use a classical computation of Riemannian geometry, about the Hessian of the *distance* (not squared!): If $k(t) = \langle \nabla^2 d(y,x)\cdot e_i(t), e_i(t)\rangle_{\gamma(t)}$, then

$$\dot k(t) + k(t)^2 + \sigma(t) \le 0, \tag{10.43}$$

where $\sigma(t)$ is the sectional curvature of the plane generated by $\dot\gamma(t)$ and $e_i(t)$ inside $T_{\gamma(t)}M$. Then we can relate $k(t)$ and $h(t)$. Indeed,

$$\nabla_x \left(\frac{d(y,x)^2}{2} \right) = d(y,x)\, \nabla_x d(y,x); \qquad \nabla_x^2 \left(\frac{d(y,x)^2}{2} \right) = d(y,x)\, \nabla_x^2 d(y,x) + \nabla_x d(x,y) \otimes \nabla_x d(x,y).$$

By applying this to the tangent vector $e_i(t)$ and using the fact that $\nabla_x d(x,y)$ at $x = \gamma(t)$ is just $\dot\gamma(t)$, we get

$$h(t) = d(y,\gamma(t))\, k(t) + \langle \dot\gamma(t), e_i(t) \rangle^2 = t\, k(t).$$

Plugging this in (10.43) results in

$$t\dot h(t) - h(t) + h(t)^2 \le -t^2\, \sigma(t). \tag{10.44}$$

From (10.44) follow the two comparison results which were used in Theorem 10.35 and Corollary 10.38:

(a) Assume that **the sectional curvatures of M are all nonnegative**. Then (10.44) forces $\dot h \le 0$, so h remains bounded above by 1 for all times. This proves that

$$\text{nonnegative sectional curvature} \implies \qquad \nabla_x^2 \left(\frac{d(x,y)^2}{2} \right) \le \mathrm{Id}_{\,T_xM}. \tag{10.45}$$

(If we think of the Hessian as a bilinear form, this is the same as $\nabla_x^2 (d(x,y)^2/2) \le g$, where g is the Riemannian metric.) Inequality (10.45) is rigorous if $d(x,y)^2/2$ is twice differentiable at x; otherwise the conclusion should be reinterpreted as

$$x \to \frac{d(x,y)^2}{2} \quad \text{is semi-concave with a modulus } \omega(r) = \frac{r^2}{2}.$$

(b) Assume now that **the sectional curvatures at point x are bounded below by $-C/d(x_0,x)^2$**, where x_0 is an arbitrary point. In this case I shall say that M is *asymptotically flat*. Then if y varies in a compact subset, we have a lower bound like $\sigma(t) \ge -C'/d(y,x)^2 = -C'/t^2$, where C' is some positive constant. So (10.44) implies

$$t\dot h(t) \le C' + h(t) - h(t)^2.$$

If $h(t)$ ever becomes strictly greater than $\overline{C} := (1/2) + (C' + 1/4)^(1/2)$, then the right-hand side becomes strictly negative; so h can never go above $\overline{C}$. The conclusion is that

$$M \text{ is asymptotically flat} \implies \qquad \forall y \in K, \quad \nabla_x^2 \left(\frac{d(x,y)^2}{2} \right) \le C(K)\, \mathrm{Id}_{\,T_xM},$$

where K is any compact subset of M. Again, at points where $d(x,y)^2/2$ is not twice differentiable, the conclusion should be reinterpreted as

$$x \to \frac{d(x,y)^2}{2} \quad \text{is semi-concave with a modulus } \omega(r) = C(K)\, \frac{r^2}{2}.$$

Example 10.45. Any compact manifold is asymptotically flat in the preceding sense. Any manifold which has been obtained from $\mathbb{R}^n$ by modification on a compact set is also asymptotically flat. The hyperbolic space $\mathbb{H}^n$ is not asymptotically flat. In fact, if y is any given point in $\mathbb{H}^n$, then the function $x \to d(y,x)^2$ is *not* uniformly semi-concave as $x \to \infty$.

Remark 10.46. The exponent 2 appearing in the definition of "asymptotic flatness" above is optimal in the sense that for any $p < 2$ it is possible to construct manifolds satisfying $\sigma_x \ge -C/d(x_0,x)^p$ and on which $d(x_0,x)^2$ is not uniformly semi-concave as a function of x.

Bibliographical Notes

The existence of solutions to the Monge problem and the differentiability of c-convex functions, for strictly superlinear convex cost functions in $\mathbb{R}^n$ (other than quadratic) was investigated by several authors, including in particular Rüschendorf [318] (formula (10.4) seems to appear there for the first time), Knott and Smith [331], Gangbo and McCann [178, 179]. In the latter reference, the authors showed how to get rid of all moment assumptions by avoiding the explicit use of Kantorovich duality. These results are reviewed in [365, Chapter 2]. Gangbo and McCann imposed some assumptions of growth and superlinearity, in particular the one described in Example 10.18.

Gangbo and McCann [179] also investigated the case of strictly concave cost functions in $\mathbb{R}^n$ (more precisely, strictly concave functions of the distance), which are probably more realistic from an economic perspective, as explained in the introduction of their paper. The main results from [179] are briefly reviewed in [365, Section 2.4]. Further numerical and theoretical analysis for non-convex cost functions in dimension 1 are considered by McCann [268], and Rüschendorf and Uckelmann [324]. Hsu showed me a very nice application of an optimal transport problem with a concave cost to a problem of maximal coupling of Brownian paths, on which he worked together with Sturm.

McCann [269] proved Theorem 10.35 when M is a compact Riemannian manifold and μ is absolutely continuous. In his paper he also mentioned the possibility to cover more general cost functions expressed in terms of the distance. Later Bernard and Buffoni [47] generalized his results to Lagrangian cost functions, and imported tools and techniques from Lagrangian systems theory (related in particular to Mather's minimization problem). Before that explicit link, several researchers, in particular Evans, Fathi and Gangbo, had become gradually aware of the strong similarities between Monge's theory on one hand, and Mather's theory on the other hand.

Feyel and Üstünel [**?**] derived theorems of unique solvability of the Monge problem in the Wiener space, when the cost is the square of the Cameron–Martin distance (or rather pseudo-distance, since it takes the value $+\infty$). Their tricky analysis goes via finite-dimensional approximations.

The use of approximate differentials as in Theorem 10.32 was initiated by Ambrosio and collaborators [15], who used it for strictly convex cost functions in $\mathbb{R}^n$. The nontrivial adaptation to Riemannian manifolds was then done by Fathi and Figalli [161], with a slightly more complicated approach than the one used in this chapter.

The tricky proof of Theorem 10.36 takes its roots in a uniqueness theorem by Alexandrov [8]. McCann [266] understood that Alexandrov's strategy could be revisited to yield the uniqueness of a cyclically monotone transport in $\mathbb{R}^n$ without the assumption of finite total cost (Corollary 10.38 in the case when $M = \mathbb{R}^n$). The tricky extension to more general cost functions on Riemannian manifolds was performed later by Figalli. The current proof of Theorem 10.36 is so complicated that the reader might prefer to have a look at [365, Section 2.3.3], where the core of McCann's proof is explained in simpler terms on the particular case $c(x, y) = |x - y|^2$.

The case when the cost function is the distance ($c(x, y) = d(x, y)$) is not covered by Theorem 10.26, nor by any of the theorems appearing in the present chapter. This case is actually much more tricky, be it in Euclidean space or on a manifold. The interested reader can consult [365, Section 2.4.6] for a brief review, as well as the research papers [342, 10, 17, 93, 357, 162, 48, 16, 130, 170]. The treatment by Bernard and Buffoni [48] is particularly appealing, for its simplicity and links to dynamical system tools.

The optimal transport problem with a distance cost function is also related to the so-called irrigation problem studied recently by various authors [49, 84], the Bouchitté-Buttazzo variational problem [67, 66], and other problems as well. In this connection, see also Pratelli [300].

To conclude, here are some remarks about some key technical ingredients used in this chapter.

Rademacher [308] proved his theorem of almost everywhere differentiability in 1918, for Lipschitz functions of two variables; this was later generalized to an arbitrary number of variables. The simple argument presented in this section seems to be due to Christensen [112]; it can also be found, up to minor variants, in modern textbooks about real analysis such as the one by Evans and Gariepy [156, pp. 81–84]. Ambrosio showed me another simple argument which uses Lebesgue's density theorem and the identification of a Lipschitz function with a function whose distributional derivative is essentially bounded.

The book by Cannarsa and Sinestrari [94] is an excellent reference for semi-concavity (or semi-convexity), and subdifferentiability (or superdifferentiability) in $\mathbb{R}^n$, as well as the links with the theory of Hamilton–Jacobi equations. It is centered on semiconcavity rather than semiconvexity, but this is just a question of convention. In that reference it is proven that a semiconcave function is locally Lipschitz in the interior of its domain [94, Theorem 2.1.7]; that a subset of $\mathbb{R}^n$ whose tangent spaces are all of dimension at most d is countably d-rectifiable [94, Theorem 4.1.6 and Corollary 4.1.9]; that a semi-concave function is differentiable on the interior of its domain, apart from a countably $(n-1)$-rectifiable set [94, Corollary 4.1.13]; etc. In particular, the proof of Theorem 10.41(i) is taken from that source. The core results in this circle of ideas and tools can be traced back to a pioneering paper by Alberti, Ambrosio and Cannarsa [5]. Following Ambrosio's advice, I used the same methods to establish Theorem 10.41(ii) in the present notes.

Apart from subdifferentiability, other notions of differentiability for non-smooth functions are discussed in [94], such as the Dini derivatives, the reachable gradients or the Clarke gradient.

The theory of approximate differentiability is developed in Federer [**?**, Section 3.1.8] (in the context of Euclidean space); see also Ambrosio, Gigli and Savarè [15, Section 5.5]. A central result is the fact that any approximately differentiable function coincides, up to a set of arbitrarily small measure, with a Lipschitz function.

Besicovich's density theorem can be found in [**?**]; it is based on Besicovich's covering lemma. This theorem is an alternative to the more classical Lebesgue density theorem (based on Vitali's covering lemma), which requires the doubling property. The price to pay for Besicovich's theorem is that it only works in $\mathbb{R}^n$ (or a Riemannian manifold, by localization) rather than on a general metric space.

The nonsmooth implicit function theorem stated in the second Appendix (Theorem 10.43) seems to be folklore in nonsmooth real analysis, although I am not aware of a precise reference other than the present notes. The core of the proof of Theorem 10.43 was explained to me by Fathi. Corollary 10.44 was discovered or rediscovered by McCann [266, Appendix], in the case where ψ and $\widetilde{\psi}$ are convex functions in $\mathbb{R}^n$.

Everything in the Third Appendix, in particular the key differential inequality (10.44), was explained to me by Gallot. The lower bound assumption on the sectional curvatures $\sigma_x \geq -C/d(x_0, x)^2$ is sufficient to get upper bounds on $\nabla_x^2 d(x,y)^2$ as y stays in a compact set, but it is *not sufficient* to get upper bounds that are uniform in both x and y. A counterexample is developed in [**?**, **?**, **?**, pp.213–214]

As pointed out to me by Ghys, the problem of finding a sufficient condition for the Hessian of $d(x,y)^2$ to be bounded above is closely related to the problem whether large

spheres $S_r(y)$ centered at y look flat at infinity, in the sense that their second fundamental form is bounded like $O(1/r)$.

11

The Jacobian equation

Transport is but a change of variables, and in many problems involving changes of variables, it is useful to write the Jacobian equation

$$f(x) = g(T(x))\, \mathcal{J}_T(x),$$

where f and g are the respective densities of the probability measures μ and ν with respect to the volume measure (in $\mathbb{R}^n$, the Lebesgue measure), and $\mathcal{J}_T(x)$ is the absolute value of the Jacobian determinant associated with T:

$$\mathcal{J}_T(x) = |\det(\nabla T(x))| = \lim_{r \to 0} \frac{\mathrm{vol}\,[T(B_r(x))]}{\mathrm{vol}\,[B_r(x)]}.$$

There are two important things that one should check before writing the Jacobian equation: First, T should be *injective* on its domain of definition; secondly, it should possess some minimal *regularity*.

So how smooth should T be for the Jacobian equation to hold true? We learn in elementary school that it is sufficient for T to be continuously differentiable, and a bit later that it is actually enough to have T Lipschitz continuous. But that degree of regularity is not always available in optimal transport! As we shall see in the Appendix of Chapter 12, the transport map T might fail to be even continuous.

There are (at least) three ways out of this situation:

(i) Only use the Jacobian equation in situations where the optimal map is smooth. As explained in Chapter 12, such situations are rare. For instance, if the cost function is the square distance on a Riemannian manifold, M then known theorems of regularity of the optimal transport apply only when M is the Euclidean space or the sphere; and moreover f and g should satisfy some conditions of strict positivity.

(ii) Only use the Jacobian equation for the optimal map between μ_{t_0} and μ_t, where $(\mu_t)_{0 \le t \le 1}$ is a compactly supported displacement interpolation, and t_0 is fixed in $(0, 1)$. Then, according to Theorem 8.5, the transport map is essentially Lipschitz. This is the strategy that I shall use in these notes.

(iii) Apply a more sophisticated theorem of change of variables, covering for instance changes of variables with bounded variation (possibly discontinuous). It is in fact sufficient that the map T be differentiable almost everywhere, or even just approximately differentiable almost everywhere, in the sense of Definition 10.2. Such a theorem is stated here without proof. The volume measure on M will be denoted just dx.

Theorem 11.1 (Jacobian equation). *Let M be a Riemannian manifold, let $f \in L^1(M)$ be a nonnegative integrable function on M, and let $T : M \to M$ be a Borel map. Define $\mu(dx) = f(x)\, dx$ and $\nu := T_\# \mu$. Assume that*

(i) There exists a measurable set $\Sigma \subset M$, such that $f = 0$ almost everywhere outside of Σ, and T is injective on Σ;

(ii) T is approximately differentiable almost everywhere on Σ, with approximate gradient $\widetilde{\nabla T}$.

Let then $\mathcal{J}_T$ be defined almost everywhere on Σ by the equation $\mathcal{J}_T(x) := |\det(\widetilde{\nabla} T(x))|$. Then ν is absolutely continuous with respect to the volume measure if and only if $\mathcal{J}_T > 0$ almost everywhere. In that case ν is concentrated on $T(\Sigma)$, and its density is determined by the equation

$$f(x) = g(T(x))\, \mathcal{J}_T(x). \tag{11.1}$$

In an informal writing:

$$\frac{d(T^{-1})_\#(g\,\mathrm{vol}\,)}{d\mathrm{vol}} = \mathcal{J}_T\,(g \circ T)\,\mathrm{vol}\,.$$

Theorem 11.1 establishes the Jacobian equation as soon as, say, the optimal transport has locally bounded variation. Indeed, in this case the map T is almost everywhere differentiable, and its gradient coincides with the absolutely continuous part of the distributional gradient $\nabla_{\mathcal{D}'} T$. The property of bounded variation is obviously satisfied for the quadratic cost in Euclidean space, since the second derivative of a convex function is a nonnegative measure.

Example 11.2. Consider two probability measures μ_0 and μ_1 on $\mathbb{R}^n$, with finite second moments; assume that μ_0 and μ_1 are absolutely continuous with respect to Lebesgue measure, with respective densities f_0 and f_1. Under those assumptions there exists a unique optimal transport map between μ_0 and μ_1, and it takes the form $T(x) = \nabla\Psi(x)$ for some convex Ψ. There is a unique displacement interpolation $(\mu_t)_{0\le t\le 1}$, and it is defined by

$$\mu_t = (T_t)_\#\mu_0, \qquad T_t(x) = (1-t)\,x + t\,T(x) = (1-t)\,x + t\,\nabla\Psi(x).$$

By Theorem 8.7, each μ_t is absolutely continuous, so let f_t be its density. The map ∇T is of locally bounded variation, and it is differentiable almost everywhere, with Jacobian matrix $\nabla T = \nabla^2\Psi$, where $\nabla^2\Psi$ is the Alexandrov Hessian (absolutely continuous part of the distributional Hessian) of Ψ. Then, it follows from Theorem 11.1 that, μ_0-almost surely,

$$f_0(x) = f_1(\nabla\Psi(x))\,\det(\nabla^2\Psi(x)).$$

Also, for any $t \in [0,1]$,

$$\begin{aligned} f_0(x) &= f_t(T_t(x))\,\det(\nabla T_t(x)) \\ &= f_t\big((1-t)x + t\nabla\Psi(x)\big)\,\det\big((1-t)I_n + t\nabla^2\Psi(x)\big), \end{aligned}$$

If now $T_{t_0\to t} = T_t \circ T_{t_0}^{-1}$ stands for the transport map between μ_{t_0} and μ_t, then the equation

$$f_{t_0}(x) = f_t(T_{t_0\to t}(x))\,\det(\nabla T_{t_0\to t}(x))$$

also holds true for $t_0 \in (0,1)$; but now this is just the theorem of change of variables for Lipschitz maps.

In the sequel of these notes, I shall be content with the following theorem of change of variables.

Theorem 11.3 (Change of variables). *Let M be a Riemannian manifold, and $c(x,y)$ a cost function deriving from a C^2 Lagrangian $L(x,v,t)$ on $TM \times [0,1]$, where L satisfies the classical assumptions of Definition 7.6, together with $\nabla_v^2 L > 0$. Let $(\mu_t)_{0\le t\le 1}$ be a displacement interpolation, such that μ_t is absolutely continuous and has density f_t. Let $t_0 \in (0,1)$, and $t \in [0,1]$; let further $T_{t_0\to t}$ be the (μ_{t_0}-a.s.) unique optimal transport from μ_{t_0} to μ_t, and let $\mathcal{J}_{t_0\to t}$ be the associated Jacobian determinant. Let F be a nonnegative measurable function on $M \times \mathbb{R}_+$ such that*

$$f_t(y) = 0 \Longrightarrow \quad F(y, f_t(y)) = 0.$$

Then,

$$\int_M F(y, f_t(y))\, dy = \int_M F\Big(T_{t_0\to t}(x),\, \frac{f_{t_0}(x)}{\mathcal{J}_{t_0\to t}(x)}\Big) \mathcal{J}_{t_0\to t}(x)\, dx.$$

Furthermore, for μ_{t_0}-almost all x, the Jacobian determinant $\mathcal{J}_{t_0\to t}(x)$ is positive for all $t \in [0,1]$.

Proof of Theorem 11.3. Let us first consider the case when $(\mu_t)_{0\le t\le 1}$ is compactly supported. Let Π be a probability measure on the set of minimizing curves, such that $\mu_t = (e_t)_\# \Pi$. Let $K_t = e_t(\mathrm{Spt}\,\Pi)$ and $K_{t_0} = e_{t_0}(\mathrm{Spt}\,\Pi)$. By Corollary 8.2, the map $\gamma_{t_0} \to \gamma_t$ is well-defined and Lipschitz for all $\gamma \in \mathrm{Spt}\,\Pi$. So $T_{t_0\to t}(\gamma_{t_0}) = \gamma_t$ is a Lipschitz map $K_{t_0} \to K_t$. Moreover, (γ_t, γ_{t_0}) is a deterministic coupling of (μ_t, μ_{t_0}), and it is optimal for the cost $c^{t_0,t}$ obtained by exchanging the variables in c^{t,t_0}. (That is, $c^{t_0,t}(x,y) = c^{t,t_0}(y,x)$.) By assumption μ_t is absolutely continuous, so Theorem 10.26 guarantees that the coupling (γ_t, γ_{t_0}) is deterministic, which amounts to say that $\gamma_{t_0} \to \gamma_t$ is injective apart from a set of zero probability.

Then we can use the change of variables formula with $g = 1_{K_t}$, $T = T_{t_0\to t}$, and we find $f(x) = \mathcal{J}_{t_0\to t}(x)$. Therefore, for any nonnegative measurable function G on M,

$$\int_{K_t} G(y)\, dy = \int_{K_t} G(y)\, d((T_{t_0\to t})_\# \mu)(y) = \int_{K_{t_0}} (G \circ T_{t_0\to t})\, f(x)\, dx = \int_{K_{t_0}} G(T_{t_0\to t}(x))\, \mathcal{J}_{t_0\to t}(x)\, dx.$$

We can apply this with $G(y) = F(y, f_t(y))$ and then replace $f_t(T_{t_0\to t}(x))$ by $f_{t_0}(x)/\mathcal{J}_{t_0\to t}(x)$; this is allowed since in the right-hand side the contribution of those x with $f_t(T_{t_0\to t}(x))$ is negligible, and $\mathcal{J}_{t_0\to t}(x) = 0$ implies (almost surely) $f_t(T_{t_0\to t}(x)) = 0$. So in the end

$$\int_{K_t} F(y, f_t(y))\, dy = \int_{K_{t_0}} F\Big(T_{t_0\to t}(x), \frac{f_{t_0}(x)}{\mathcal{J}_{t_0\to t}(x)}\Big)\, \mathcal{J}_{t_0\to t}(x)\, dx.$$

Since $f_t(y) = 0$ almost surely outside of K_t and $f_{t_0}(x) = 0$ almost surely outside of K_{t_0}, these two integrals can be extended to the whole of M.

Now it remains to generalize this to the case when Π is not compactly supported. (The reader may skip this bit at first reading.) In this case we let $(K_\ell)_{\ell\in\mathbb{N}}$ be an nondecreasing sequence of compact sets, such that $\Pi[\cup K_\ell] = 1$. For ℓ large enough, $\Pi[K_\ell] > 0$, so we can consider the restriction Π_ℓ of Π to K_ℓ. Then we let $K_{t,\ell}$ and $K_{t_0,\ell}$ be the images of K_ℓ by e_t and e_{t_0}, and of course $\mu_{t,\ell} = (e_t)_\# \Pi_\ell$, $\mu_{t_0,\ell} = (e_{t_0})_\# \Pi_\ell$. Since μ_t and μ_{t_0} are absolutely continuous, so are $\mu_{t,\ell}$ and $\mu_{t_0,\ell}$, so let $f_{t,\ell}$ and $f_{t_0,\ell}$ be their respective densities. The optimal map $T_{t_0\to t,\ell}$ for the transport problem between $\mu_{t_0,\ell}$ and $\mu_{t,\ell}$ is obtained as before by the map $\gamma_{t_0} \to \gamma_t$, so this is actually the restriction of $T_{t_0\to t}$ to $K_{t_0,\ell}$. So we have the Jacobian equation

$$f_{t_0,\ell}(x) = f_{t,\ell}(T_{t_0\to t}(x))\, \mathcal{J}_{t_0\to t}(x). \tag{11.2}$$

Note that the Jacobian determinant does not depend on ℓ. This equation holds true almost surely for $x \in K_{\ell'}$, as soon as $\ell' \le \ell$, so we may pass to the limit as $\ell \to \infty$ to get the Jacobian equation

$$f_{t_0}(x) = f_t(T_{t_0 \to t}(x))\, \mathcal{J}_{t_0 \to t}(x). \tag{11.3}$$

This equation holds true almost surely on $K_{\ell'}$, for each ℓ', so it also holds true almost surely.

Next, for any nonnegative measurable function G, by monotone convergence and the first part of the proof one has

$$\int_{UK_{t,\ell}} G(y)\, dy = \lim_{\ell \to \infty} \int_{K_{t,\ell}} G(y)\, dy = \lim_{\ell \to \infty} \int_{K_{t_0,\ell}} G(T_{t_0 \to t}(x))\, \mathcal{J}_{t_0 \to t}(x)\, dx$$
$$= \int_{UK_{\ell,t_0}} G(T_{t_0 \to t}(x))\, \mathcal{J}_{t_0 \to t}(x)\, dx.$$

The conclusion follows as before by choosing $G(y) = F(y, f_t(x))$ and using the Jacobian equation (11.3), then extending the integrals to the whole of M.

It remains to prove the assertion about $\mathcal{J}_{t_0 \to t}(x)$ being positive *for all values of* $t \in [0,1]$, and not just for almost all values of t. The transport map $T_{t_0 \to t}$ can be written $\gamma(t_0) \to \gamma(t)$, where γ is a minimizing curve determined uniquely by $\gamma(t_0)$. Since γ is minimizing, we know (Recall Problem 8.9) that the map $(\gamma_0, \dot\gamma_0) \to (\gamma_0, \gamma_{t_0})$ is locally invertible. So $T_{t_0 \to t}$ can be written as the composition of the maps $F_1 : \gamma(t_0) \to (\gamma(0), \gamma(t_0))$, $F_2 : (\gamma(0), \gamma(t_0)) \to (\gamma(0), \dot\gamma(0))$ and $F_3 : (\gamma(0), \dot\gamma(0)) \to \gamma(t)$. Both F_2 and F_3 have positive Jacobian determinant, at least if $t < 1$; so if x is chosen in such a way that F_1 has positive Jacobian determinant at x, then also $T_{t_0 \to t} = F_3 \circ F_2 \circ F_1$ will have positive Jacobian determinant at x. □

Bibliographical Notes

Theorem 11.1 can be obtained (in $\mathbb{R}^n$) by combining Lemma 5.5.3 in [15] with Theorem 3.83 in [13].

In the context of optimal transport, the change of variables formula (11.1) was proven by McCann [267]. His argument is based on Lebesgue's density theory, and takes advantage of *Alexandrov's theorem* (alluded to in this chapter): *A convex (or semi-convex) function admits a Taylor expansion at order 2 at almost each x in its domain of definition.* In short, a convex function is twice differentiable almost everywhere with respect to Lebesgue measure. Since the gradient of a convex function has locally bounded variation, Alexandrov's theorem can be seen essentially as a particular case of the theorem of approximate differentiability of functions with bounded variation. Strictly speaking, this is not exactly true, since the second differentiability in Alexandrov's theorem is not stated in terms of differentiability of the gradient, but in terms of the existence of a second-order Taylor expansion.

Altogether McCann's proof is somewhat easier than the proof of Proposition 11.1, which does not assume that T is the gradient of a convex function. His argument is reproduced in [365, Theorem 4.8]; to complete the proof one should complement this with Alexandrov's theorem [7] (for "modern" proofs see [156, pp. 241–245], [4, Theorem 7.10], or the Appendix of Chapter 14).

Along with Cordero-Erausquin and Schmuckenschläger, McCann later generalized his result to the case of Riemannian manifolds [118]. Modulo certain complications, the proof

basically follows the same pattern as in $\mathbb{R}^n$. Then Cordero-Erausquin [116] treated the case of strictly convex cost functions in $\mathbb{R}^n$ in a similar way.

Later Ambrosio pointed out that those results could be retrieved within the general framework of push-forward by approximately differentiable mappings. This point of view has the disadvantage to involve more subtle arguments, but the advantage to show that it is not a special feature of optimal transport. It also applies to nonsmooth cost functions such as $|x - y|^p$. In fact it covers general strictly convex cost of the form $c(x - y)$ as soon as c has superlinear growth, is C^1 everywhere and C^2 out of the origin. A more precise discussion of these subtle issues can be found in [15, Section 6.2.1].

It is a general feature of optimal transport with strictly convex cost in $\mathbb{R}^n$ that if T stands for the optimal transport map, then the matrix ∇T, even if not necessarily nonnegative symmetric, is diagonalizable with nonnegative eigenvalues; see Cordero-Erausquin [116] and Ambrosio, Gigli and Savaré [15, Section 6.2]. From an Eulerian perspective, that diagonalizability property was already understood by Otto [287]. I don't know if there is an analogue on Riemannian manifolds.

A remarkable contribution by Cabré [85] uses the Jacobian properties of $d^2/2$-convex functions to investigate qualitative properties of elliptic equations (Liouville theorem, Alexandrov-Bakelman-Pucci estimates, Krylov-Safonov-Harnack inequality) on Riemannian manifolds with nonnegative sectional curvature.

12

Smoothness

If we are going to use optimal transport in practical computations, it might certainly help to have information about its smoothness. So what regularity can be expected on the optimal transport T?? What characterizes T is the existence of a ψ such that (10.17) (or (10.20)) holds true; so it is natural to search for a closed equation on ψ.

To guess the equation, let us work formally without being too demanding about regularity issues, and also let us assume that we work in $\mathbb{R}^n$. As we shall see, even in that case we shall arrive at a rather negative conclusion. Let $\mu(dx) = f(x)\,dx$ and $\nu(dy) = g(y)\,dy$ be two absolutely continuous probability measures, let $c(x, y)$ be a smooth cost function, and let T be a Monge transport. The differentiation of (10.17) with respect to x (once again) leads to

$$\nabla^2\psi(x) + \nabla^2_x c(x, T(x)) + \nabla^2_{xy} c(x, T(x)) \cdot \nabla T(x) = 0,$$

which can be rewritten

$$\nabla^2_{xx} c(x, T(x)) + \nabla^2\psi(x) = -\nabla^2_{xy} c(x, T(x)) \cdot \nabla T(x). \tag{12.1}$$

The expression on the left-hand side is the Hessian of the function $c(x', T(x)) + \psi(x')$, considered as a function of x' and then evaluated at x. Since this function is minimum for $x' = x$, its Hessian is nonnegative, so the left-hand side of (12.1) is a nonnegative symmetric matrix; in particular its determinant is nonnegative. Take absolute values of determinants on both sides of (12.1):

$$\det\Big(\nabla^2_{xx} c(x, T(x)) + \nabla^2\psi(x)\Big) = \big|\det \nabla^2_{xy} c(x, T(x))\big|\,|\det(\nabla T(x))|.$$

Then the Jacobian determinant in the right-hand side can be replaced by $f(x)/g(T(x))$, and we arrive at

$$\det\Big(\nabla^2_{xx} c(x, T(x)) + \nabla^2\psi(x)\Big) = \big|\det \nabla^2_{xy} c(x, T(x))\big|\,\frac{f(x)}{g(T(x))}. \tag{12.2}$$

This becomes a closed equation on ψ in terms of f and g, if one recalls from (10.17) that

$$T(x) = (\nabla_x c)^{-1}\big(x, -\nabla\psi(x)\big), \tag{12.3}$$

where the inverse is with respect to the y variable.

Unfortunately there is no simplification to expect, except in special cases. The most important of them is the quadratic cost function, or equivalently $c(x, y) = -x \cdot y$ in $\mathbb{R}^n$. Then (12.2)–(12.3) reduces to

$$\det \nabla^2 \psi(x) = \frac{f(x)}{g(\nabla \psi(x))}. \tag{12.4}$$

This partial differential equation is an instance of the *Monge–Ampère equation*, and the regularity of its solutions has been studied by several authors. Here is a typical result: If f and g are bounded below and smooth in the interior of their respective support, *and the support of g is convex*, then the function ψ is as smooth as can be hoped — that is, basically two degrees of regularity better than f and g themselves. This statement should be taken cautiously; see the bibliographical notes for details.

At this point we may have the feeling that partial differential equations theory will help our task quite a bit by providing regularity results for the optimal map in the Monge—Kantorovich problem, at least if we rule out cases where the map is trivially discontinuous (for instance if the support of the initial measure μ is connected, while the support of the final measure ν is not). To a certain extent, this is true; for instance, one has the following result: If f and g are bounded below and smooth in the interior of their respective support, *and the support of g is convex*, then the function ψ in (12.2) is as smooth as can be hoped — that is, basically two degrees of regularity better than f and g themselves. (See the bibliographical notes for more precise formulations.)

However, the truth is that in many cases of interest, the optimal transport will not be smooth, as I shall illustrate by some counterexamples. As a temporary conclusion: If we want to use optimal transport in rather general situations, we'd better find ways to do without regularity. Actually, it is one of the striking facts in the theory of optimal transport that it can be pushed very far with almost no regularity available.

Caffarelli's counterexample

Caffarelli understood that regularity results for (12.2) in $\mathbb{R}^n$ cannot be obtained unless one adds an assumption of *convexity of the target support*. Without such an assumption, the optimal transport may very well be discontinuous, as the next counterexample shows.

Theorem 12.1 (An example of discontinuous optimal transport). *There are smooth compactly supported probability densities f and g on $\mathbb{R}^n$, such that the supports of f and g are smooth and connected, f and g are strictly positive in the interior of their respective supports, and yet the optimal transport between $\mu(dx) = f(x)\,dx$ and $\nu(dy) = g(y)\,dy$ is discontinuous.*

Proof. Let f be the indicator function of the unit ball B in $\mathbb{R}^2$ (normalized to be a probability measure), and let $g = g_\varepsilon$ be the (normalized) indicator function of a set C_ε obtained by first separating the ball into two halves B_1 and B_2 (say with distance 2); and then building a thin bridge between those two halves, of width $O(\varepsilon)$. Let also g be the normalized indicator function of $B_1 \cup B_2$: this is the limit of g_ε as $\varepsilon \downarrow 0$. It is not difficult to see that g (identified with a probability measure) can be obtained from f by a continuous deterministic transport (after all, one can deform B continuously into C_ε; just think that you are playing with clay, then it is possible to massage the ball into C_ε, without tearing off). However, we shall see here that for ε small enough, the optimal transport *cannot be continuous*.

The proof will rest on the stability of optimal transport: If T is the unique optimal transport between μ and ν, and T_ε is an optimal transport betwen μ and ν_ε, then T_ε converges to T in μ-probability as $\varepsilon \downarrow 0$ (Corollary 5.20).

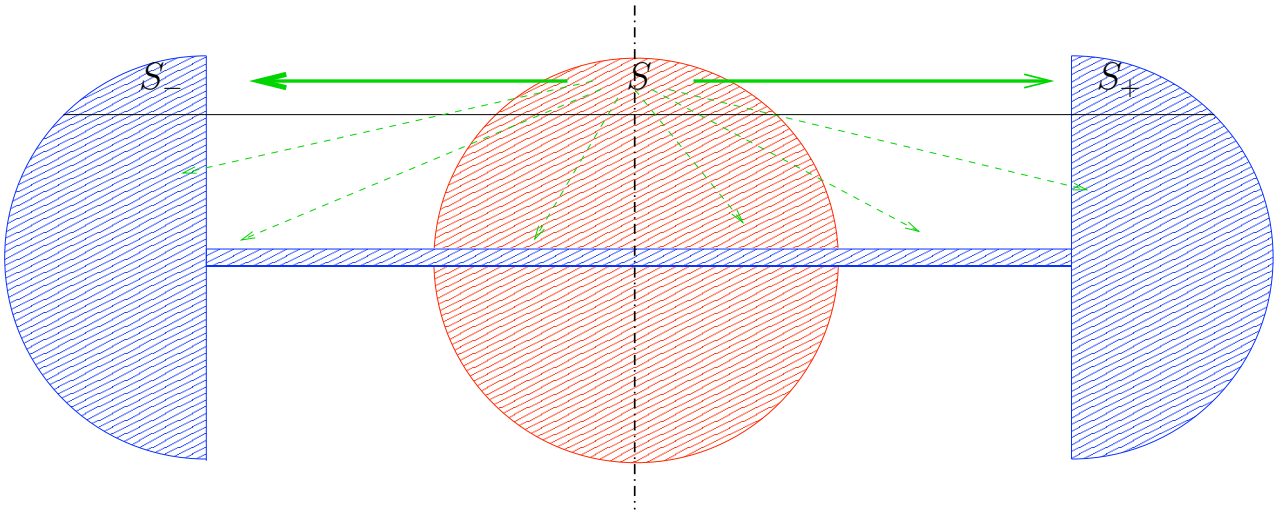


Fig. 12.1. Principle behind Caffarelli's counterexample. The optimal transport from the ball to the "dumbbells" has to be discontinuous, and in effect splits the upper region S into the upper left and upper right regions S_- and S_+. Otherwise, there should be some transport along the dashed lines, but for some lines this would contradict monotonicity.

In the present case, choosing $\mu(dx) = f(x)\,dx$ and $\nu(dy) = g(y)\,dy$, and then choosing the cost function to be $c(x,y) = |x-y|^2$, it is easy to figure out that the unique optimal transport T is the one that sends (x,y) to $(x-1,y)$ if $x<0$, and to $(x+1,y)$ if $x>0$.

Let now S, S_+ and S_- be as on Figure 12.1. From the convergence in probability, it follows that, for ε small enough, a large fraction (say 0.99) of the mass in S has to go to S_- (if it lies on the left) or to S_+ (if it lies on the right). Since the continuous image of a connected set is itself connected, there have to be some points in $T_\varepsilon(S)$ that form a path going from S_- to S_+; and so there are some points x such that $T_\varepsilon(x) - x$ is pointing downwards and to the left, say with a 45^o angle. Let x be one such point. From the convergence in probability again, many of the neighbors of x have to be transported to S_-, with nearly horizontal displacements $T(\widetilde{x}) - \widetilde{x}$. It is not difficult to check that one of these $\widetilde{x}$ will contradict the fact that $\big\langle x-\widetilde{x}, T(x)-T(\widetilde{x})\big\rangle$ should always be nonnegative.

The conclusion is that when ε is small enough, the optimal map T_ε is discontinuous.

The maps f and g in this example are extremely smooth (in fact constant!) in the interior of their support, but they are not smooth as maps defined on $\mathbb{R}^n$. If one wants to produce a similar construction with functions that are smooth on $\mathbb{R}^n$, this is easy: Regularize f, split it in two halves again, add a thin bridge, and make a very slight regularization of the resulting function g_ε. Then again the optimal transport will be discontinuous for ε small enough. □

Loeper's counterexample

Loeper (building on previous work by Ma, Trudinger and Wang) understood that the continuity of optimal transport in a general Riemannian setting could be prevented by some geometric obstructions.

Theorem 12.2 (A further example of discontinuous optimal transport). *There is a smooth compact Riemannian surface S, and there are two smooth positive probability densities f and g on M, such that the optimal transport between $\mu(dx) = f(x)\,dx$ and*

$\nu(dy) = g(y)\,dy$, with a cost function equal to the square of the geodesic distance on S, is discontinuous.

Remark 12.3. Loeper's results are much more precise: He shows that this phenomenon (that is, the discontinuity of the optimal transport between two smooth positive densities) occurs on any Riemannian manifold admitting a *negative sectional curvature* at some point. In fact, smoothness of optimal transport requires more than nonnegativity of the sectional curvatures: there is a more stringent necessary (and almost sufficient) condition, expressed in terms of derivatives of the metric up to order 4. See the bibliographic notes for more details.

One of the key ingredients in the proof of Theorem 12.2 is the following elementary lemma:

Lemma 12.4. *Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be any two metric probability spaces, and let T be a continuous map $\mathcal{X} \to \mathcal{Y}$, and let $\pi = (\mathrm{Id}\,, T)_\# \mu$ be the associated transport map. Then, for each $x \in \mathrm{Spt}(\mu)$, the couple $(x, T(x))$ belongs to the support of π.*

Proof of Lemma 12.4. Let x and $\varepsilon > 0$ be given. By continuity of T, there is $\delta > 0$ such that $T(B_\delta(x)) \subset B_\varepsilon(T(x))$. Without loss of generality, $\delta \le \varepsilon$. Then

$$\begin{aligned}\pi\big[B_\varepsilon(x) \times B_\varepsilon(T(x))\big] &= \mu\big[\big\{z \in X;\ z \in B_\varepsilon(x) \text{ and } T(z) \in B_\varepsilon(T(x))\big\}\big] \\ &\ge \mu[B_\varepsilon(x) \cap B_\delta(x)] = \mu[B_\delta(x)] > 0.\end{aligned}$$

Since ε is arbitrarily small, this shows that π attributes positive measure to any neighborhood of $(x, T(x))$, which proves the claim. □

Proof of Theorem 12.2. Let S be a compact surface in $\mathbb{R}^3$ with the following properties: (a) S is invariant under the symmetries $x \to -x$, $y \to -y$; (b) S crosses the axis $(x, y) = (0, 0)$ at exactly two points, namely $O = (0, 0, 0)$ and O'; (c) S coincides in a an open ball $B(O, r)$ with the "horse saddle" $(z = x^2 - y^2)$. (Think of S as a small piece of the horse saddle which has been completed into a closed surface.)

Let $A_+ = (x_0, 0, 1 + x_0^2)$, $A_- = (-x_0, 0, 1 + x_0^2)$, and similarly let $B_+ = (0, y_0, 1 - y_0^2)$, $B_- = (0, -y_0, 1 - y_0^2)$; in the sequel the symbol $A_\pm$ will stand for "either A_+ or A_-", etc.

If x_0 and y_0 are small enough then A_+, A_-, B_+, B_- belong to a neighborhood of O where S has strictly negative curvature, and the unique geodesic joining O to $A_\pm$ (resp. $B_\pm$) satisfies the equation $(y = 0)$ (resp. $x = 0$); then the lines $(O, A_\pm)$ and $(O, B_\pm)$ are orthogonal at O. Since we are on a negatively curved surface, Pythagore's identity in a triangle with a square angle is modified in favor of the diagonal, so

$$d(N, A_\pm)^2 + d(N, B_\pm)^2 < d(A_\pm, B_\pm)^2.$$

By continuity, there is $\varepsilon_0 > 0$ small enough that the balls $B(A_+, \varepsilon_0)$, $B(A_-, \varepsilon_0)$, $B(B_+, \varepsilon_0)$ and $B(B_-, \varepsilon_0)$ are all disjoint and satisfy

$$[x \in B(A_+, \varepsilon_0) \cup B(A_-, \varepsilon_0), \quad y \in B(B_+, \varepsilon_0) \cup B(B_-, \varepsilon_0)] \implies \quad d(O, x)^2 + d(O, y)^2 < d(x, y)^2. \tag{12.5}$$

Next let f and g be smooth probability densities on M, even in x and y, such that

$$\int_{B(A_+, \varepsilon_0) \cup B(A_-, \varepsilon_0)} f(x)\,dx > \frac{1}{2}; \qquad \int_{B(B_+, \varepsilon_0) \cup B(B_-, \varepsilon_0)} g(y)\,dy > \frac{1}{2}. \tag{12.6}$$

Let $\mu(dx) = f(x)\,dx$, $\nu(dy) = g(y)\,dy$, let T be the unique optimal transport between the measures μ and ν (for the cost function $c(x,y) = d(x,y)^2$), and let $\widetilde{T}$ be the optimal transport between ν and μ. (T and $\widetilde{T}$ are inverse of each other, at least in a measure-theoretical sense.) I claim that either T or $\widetilde{T}$ is discontinuous.

Indeed, suppose to the contrary that both T and $\widetilde{T}$ are continuous. We shall first see that necessarily $T(O) = O$. (The reasoning is by symmetry and elementary topological arguments; skip this bit if you believe it.) Since the problem is symmetric with respect to $x \to -x$ and $y \to -y$, and since there is uniqueness of the optimal transport, T maps O into a point that is invariant under these two transforms, that is either O or O'. Suppose that $T(O) = O'$. Let U be a neighborhood of O, which does not contain O'. For any $s > 0$, by continuity of T, there is a small ball $B(O, r') \subset U$ such that $T(B(O, r')) \subset B(O', s)$; since $\nu[T(B(O, r'))] = \mu[T^{-1}(T(B(O, r')))] \geq \mu[B(O, r')] > 0$, it follows that the set $T(U)$ has positive density at O'. Assume now that $T(O') = O'$; and take a neighborhood U' of O' such that U and U' lie a positive distance away of each other. By the same reasoning as before, $T(U')$ has positive density at O'. Then in any arbitrarily small neighborhood of O' there is a set of positive measure whose image by $\widetilde{T}$ has to belong to U, and a set of positive measure whose image by $\widetilde{T}$ has to belong to U'. This contradicts the continuity of $\widetilde{T}$ at O'. So necessarily $T(O') = O$; but then the two points (O, O') and (O', O) belong to the support of the optimal plan associated to T, which trivially contradict the cyclical monotonicity since $d(O, O')^2 + d(O', O)^2 > d(O, O)^2 + d(O', O')^2 = 0$. The conclusion is that $T(O) = O$; by Lemma 12.4, (O, O) belongs to the support of π.

Next, (12.6) implies that there is some transfer of mass from either $B(A_+, \varepsilon_0) \cup B(A_-, \varepsilon_0)$ to $B(B_+, \varepsilon_0) \cup B(B_-, \varepsilon_0)$; in other words, we can find, in the support of the optimal transport, some (x, y) with $x \in B(A_+, \varepsilon_0) \cup B(A_-, \varepsilon_0)$ and $y \in B(B_+, \varepsilon_0) \cup B(B_-, \varepsilon_0)$. From the previous step we know that (O, O) also lies in that support; then by c-monotonicity,

$$d(x, y)^2 + d(O, O)^2 \leq d(x, O)^2 + d(y, O)^2;$$

but this contradicts (12.5). The proof is complete. □

Open Problem 12.5. *Let f, g be two smooth positive probability measures on a compact Riemannian manifold with negative sectional curvature somewhere, and let T be the optimal transport map. Is it a priori more regular than an arbitrary BV map, and in which sense? Can one describe its singularities? Do discontinuities typically occur along smooth curves, or along a possibly fractal, intricate geometry?*

Bibliographical Notes

The modern mathematical theory of the Monge–Ampère equation was pioneered by Alexandrov [8, 9] and Pogorelov [298, 299]. A modern account can be found in the recent book by Gutiérrez [202].

The application of the theory of the Monge–Ampère equation to the problem of optimal transport was achieved independently by Caffarelli [88, 89, 90] and Urbas [360, 359], using quite different and rather sophisticated techniques. The main results can be roughly summarized by saying that if f and g are $C^{k,\alpha}$ and g is locally bounded below on its support, which is assumed to be convex, then the transport potential ψ is of regularity $C^{k+2,\alpha}$ (so the transport map is $C^{k+1,\alpha}$). Densities that are positive on the whole of $\mathbb{R}^n$ are briefly discussed in [6, Appendix]. There is also a theorem of $W^{2,p}$ (Sobolev) regularity for all p

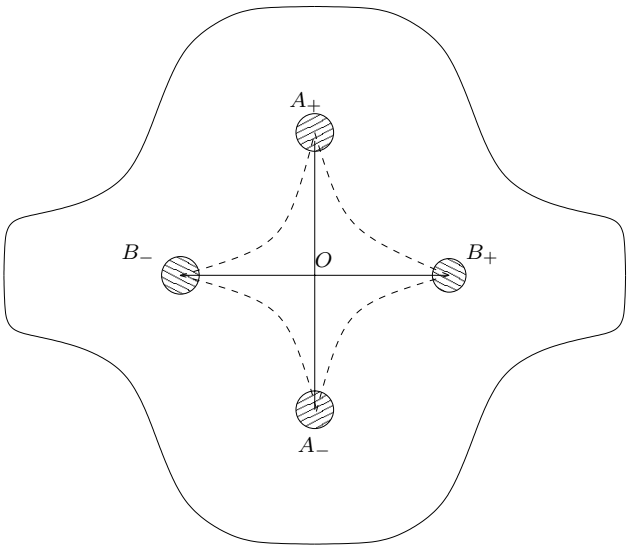


Fig. 12.2. Principle behind Loeper's counterexample. This is the surface S, immersed in $\mathbb{R}^3$, "viewed from above". By symmetry, O has to stay in place. Because most of the initial mass is close to A_+ and A_-, and most of the final mass is close to B_+ and B_-, at least some mass has to move from one of the A-balls to one of the B-balls. But then, because of the modified (negative curvature) Pythagore inequality, it is more efficient to replace the transport scheme ($A \to B$, $O \to O$), by ($A \to O$, $O \to B$).

under the assumption that f and g are continuous [87]. All of this is for the quadratic cost in $\mathbb{R}^n$.

Feyel and Üstünel [**?**] studied the infinite-dimensional Monge–Ampère equation induced by optimal transport with quadratic cost on the Wiener space.

Caffarelli's counterexample appears in Caffarelli [89], where it is used to prove that the "Hessian measure" (a generalized formulation of the Hessian determinant) cannot be absolutely continuous if the bridge is thin enough.

More general (smooth) cost functions were addressed only recently, with some pioneering works by Ma, Trudinger and Wang [250] on one hand (C^2 regularity), and Loeper [245] on the other hand ($C^{1,\alpha}$ regularity). The method in [250] takes its roots in an older paper by Wang [**?**, **?**] on the so-called antenna reflector problem, which is an optimal transport problem with cost $c(x, y) = -\log(1 - x \cdot y)$, see [**?**] or [250, Section 7.2]. The regularity theory was further developed in [**?**, 246, **?**].

It was first suggested in [250] that the regularity of solutions to the Monge–Ampère equation was closely related to Assumption (C) stated in Chapter 9. This was very much clarified in Loeper's recent work [246]. The counterexample which I discussed in these notes is a very particular case of Loeper's results; it is simple enough that one can prove the discontinuity by much simpler arguments than in [246].

To conclude these notes, I shall try to convey a crude idea of Loeper's results. For that I will need the key concept of *c-segment*: If y_0 and y_1 are two given points in $\partial_c \psi(x)$, then the c-segment $[y_0, y_1]_x$, joining y_0 to y_1, is the set of all

$$y_\theta := -(\nabla_x c)^{-1}(x, -(1-\theta)\nabla_x c(x, y_0) - \theta \nabla_x c(x, y_1)); \qquad \theta \in [0, 1].$$

Note that $-\nabla_x c(x, y_0)$ and $\nabla_x c(x, y_1)$ both belong to $\nabla^- \psi(x)$ (because $z \to \psi(z) + c(z, y_0)$ and $z \to \psi(z) + c(z, y_1)$ are minimal at $z = x$), so by convexity of the subdifferential, also $-\nabla_x c(x, y_\theta) = -(1-\theta)\nabla_x c(x, y_0) - \theta \nabla_x c(x, y_1) \in \nabla^- \psi(x)$. So $z \to \psi(z) + c(z, y_\theta)$ automatically admits a *critical point* at $z = x$; but there is no reason why this should be a minimum.

Loeper studied the regularity of optimal transport under several assumptions, that can *roughly* be formulated as follows:

(a) For any c-convex ψ and any x, the c-subdifferential $\partial_c\psi(x)$ is c-convex; that is, if y_0 and y_1 belong to $\partial_c\psi(x)$, then the c-segment $[y_0, y_1]_x$ is entirely contained in $\partial_c\psi(x)$;

(b) For any x, and for any two unit vectors ξ and ν with $\xi \perp \nu$,

$$\nabla^2_{p_\nu} \nabla^2_{x_\xi} c(x, (\nabla_x c)^{-1}(x,p)) \le 0; \tag{12.7}$$

(c) c-convex functions of class C^1 are dense (for the topology of local uniform convergence) in the set of all c-convex functions.

(d) Take any four points x_0, x_1, y_0, y_1, define the c-convex function

$$\psi(x) := \max\big(-c(x,y_0) - c(x_0,y_0),\ -c(x,y_1) - c(x_0,y_1)\big);$$

and let y_θ be any element of the c-segment $[y_0, y_1]_{x_0}$. Then the function

$$x \to \quad \psi(x) + c(x, y_\theta)$$

admits a local maximum at x_0.

Then essentially Loeper establishes the equivalence between (a), (b), (c) and (d), and shows that this condition is mandatory to have a regularity theory. If it is not fulfilled, one can find two C^∞ *positive* densities for which the solution of the associated "Monge–Ampère" equation (12.2) is not even C^1.

Examples of cost functions satisfying (b) are $c(x,y) = (1 + |x-y|^2)^{p/2}$ for $1 < p < 2$. Loeper also discusses the following reinforcement of (b):

(b') There is a positive constant C_0 such that for any x, and for any two unit vectors ξ and ν with $\xi \perp \nu$,

$$\nabla^2_{p_\nu} \nabla^2_{x_\xi} c(x, (\nabla_x c)^{-1}(x,p)) \le -C_0 \tag{12.8}$$

When Assumption (b') holds then one can develop an excellent regularity theory for (12.2), which is even slightly better than the regularity theory in Euclidean space for the "standard" Monge–Ampère equation (12.4). According to Loeper's results, this is the case for the square distance cost function on the Riemannian sphere. Actually, on the sphere there is an additional difficulty, namely the nonsmoothness of the distance function. But as shown first by Delanoë and Loeper [132], and then more precisely by Loeper [246], one can take advantage of the symmetries of the sphere to control the distance to the cut locus from below; then everything works as if the cost function were smooth. (This might be a general feature, but Loeper's proof seems to work only for the sphere.)

Of course, Condition (a) implies the connectedness of $\partial_c\psi(x)$, Assumption (C)in Chapter 9. It might even be that both conditions are equivalent.

13

Qualitative picture

This chapter is devoted to a recap of the whole picture of optimal transport on a smooth Riemannian manifold M. For simplicity I shall not try to impose the most general assumptions.

Recap

Let M be a Riemannian manifold, $L(x, v, t)$ a Lagrangian function on $TM \times [0, 1]$, satisfying the classical assumptions of Definition 7.6, together with $\nabla_v^2 L > 0$. Let $c : M \times M \to \mathbb{R}$ be the induced cost function:

$$c(x, y) = \inf\Big\{ \int_0^1 L(\gamma_t, \dot\gamma_t, t)\, dt; \quad \gamma_0 = x,\ \gamma_1 = y \Big\}.$$

More generally, define

$$c^{s,t}(x, y) = \inf\Big\{ \int_s^t L(\gamma_\tau, \dot\gamma_\tau, \tau)\, d\tau; \quad \gamma_s = x,\ \gamma_t = y \Big\}.$$

So $c^{s,t}(x, y)$ is the optimal cost to go from point x at time s, to point y at time t.

I shall consider three cases: (i) $L(x, v, t)$ arbitrary on a compact manifold; (ii) $L(x, v, t) = |v|^2/2$ on a complete manifold (so the cost is $d^2/2$, where d is the distance); (iii) $L(x, v, t) = |v|^2/2$ in $\mathbb{R}^n$ (so the cost is $|x - y|^2/2$). In all the sequel, I denote by μ_0 the initial probability measure, and by μ_1 the final one. When I say "absolutely continuous" or "singular" this is in reference with the volume measure on the manifold (Lebesgue measure in $\mathbb{R}^n$).

Recall that a *generalized optimal coupling* is a c-cyclically monotone coupling, that might be non optimal if the total cost is infinite. By analogy, I shall say that a *generalized displacement interpolation* is a path $(\mu_t)_{0 \le t \le 1}$ valued in the space of probability measures, such that $\mu_t = \text{law}\,(\gamma_t)$ and γ is a random minimizing curve such that (γ_0, γ_1) is a generalized optimal coupling. These notions are interesting only when the total cost between μ_0 and μ_1 is infinite.

By gathering the results from the previous chapters, we obtain the following information:

1. There always exists
- an optimal coupling (or generalized optimal coupling) (x_0, x_1), with law π;
- a displacement interpolation (or generalized displacement interpolation) $(\mu_t)_{0 \le t \le 1}$;
- a random minimizing curve γ with law Π;

such that law $(\gamma_t) = \mu_t$, and law $(\gamma_0, \gamma_1) = \pi$. Each curve γ is a solution of the Euler-Lagrange equation

$$\frac{d}{dt}\nabla_v L(\gamma_t, \dot\gamma_t, t) = \nabla_x L(\gamma_t, \dot\gamma_t, t). \tag{13.1}$$

In the case of a quadratic Lagrangian, this equation reduces to

$$\frac{d^2\gamma_t}{dt^2} = 0,$$

so trajectories are just geodesics, or straight lines in $\mathbb{R}^n$. Two trajectories in the support of $\mathit{\Pi}$ may intersect at time $t = 0$ or $t = 1$, but never at intermediate times.

2. If either μ_0 or μ_1 is absolutely continuous, then so is μ_t, for all $t \in (0, 1)$.

3. If μ_0 is absolutely continuous, then the optimal coupling (x_0, x_1) is unique (in law), deterministic ($x_1 = T(x_0)$) and characterized by the equation

$$\nabla\psi(x_0) = -\nabla_x c(x_0, x_1) = \nabla_v L(x_0, \dot\gamma_0, 0), \tag{13.2}$$

where $(\gamma_t)_{0\le t\le 1}$ is a minimizing curve joining $\gamma_0 = x_0$ to $\gamma_1 = x_1$ (it is part of the theorem that this curve is almost surely unique), and ψ is a c-convex function, that is, it can be written as

$$\psi(x) = \sup_{y\in M} \big[\phi(y) - c(x, y)\big]$$

for some non-trivial (i.e. not identically $-\infty$, and never $+\infty$) function ϕ. In case (ii), if nothing is known about the behavior of the distance function at infinity, then the gradient ∇ in (13.2) should be replaced by an approximate gradient $\widetilde{\nabla}$.

4. Under the same assumptions, the (generalized) displacement interpolation $(\mu_t)_{0\le t\le 1}$ is unique. This follows from the almost sure uniqueness of the minimizing curve joining γ_0 to γ_1, where (γ_0, γ_1) is the optimal coupling. (Corollary 7.21 applies when the total cost is finite; but even if the total cost is infinite, we can apply a reasoning similar to the one in Corollary 7.21.)

5. Without loss of generality, one might assume that

$$\phi(y) = \inf_{x\in M} \big[\psi(x) + c(x, y)\big]$$

(these are true supremum and true infimum, not just up to a negligible set). Moreover, one can assume without loss of generality

$$\forall x, y \in M, \qquad \phi(y) - \psi(x) \le c(x, y)$$

and

$$\phi(x_1) - \psi(x_0) = c(x_0, x_1) \qquad \text{almost surely.}$$

6. It is still possible that two minimizing curves meet at time $t = 0$ or $t = 1$, but this event may occur only on a very small set, of dimension at most $n - 1$.

7. All of the above remains true when one replaces μ_0 at time 0 by μ_t at time t, with obvious changes of notation (e.g. replace $c = c^{0,1}$ by $c^{t,1}$); the function ϕ is unchanged, but now ψ should be changed into ψ_t defined by

$$\psi_t(y) = \inf_{x\in M} \big[\psi_0(x) + c^{0,t}(x, y)\big]. \tag{13.3}$$

This ψ_t is a (viscosity) solution of the forward Hamilton–Jacobi equation

$$\partial_t \psi_t + L^*(x, \nabla \psi_t(x), t) = 0.$$

8. The equation for the optimal transport T_t between μ_0 and μ_t is as follows: $T_t(x) =$ solution at time t of the Euler-Lagrange equation starting from x with velocity

$$v_0(x) = \big(\nabla_v L(x, \cdot, 0)\big)^{-1}(\nabla \psi(x)). \tag{13.4}$$

In particular,

- for the quadratic cost on a Riemannian manifold M, $T_t(x) = \exp_x(t \nabla \psi(x))$: To obtain T_t, flow for time t along a geodesic starting at x with velocity $\nabla \psi(x)$ ($\widetilde{\nabla} \psi(x)$ if nothing is known about the behavior of M at infinity);

- for the quadratic cost in $\mathbb{R}^n$, $T_t(x) = (1-t)x + t \nabla \Psi(x)$, where $\Psi(x) = |x|^2/2 + \psi(x)$ defines a lower semi-continuous convex function in the usual sense. In particular, the optimal transport from μ_0 to μ_1 is a gradient of convex function, and this property characterizes it uniquely among all admissible transports.

Simple as they may seem by now, these statements summarize years of research. If the reader has well understood them, then he or she is ready to go on with the rest of this course. The picture is not really complete and some questions remain open, such as the following

Open Problem 13.1. *If the initial and final densities, ρ_0 and ρ_1, are positive everywhere, does it follow that the intermediate densities ρ_t are also positive? Otherwise, can one identify simple sufficient conditions for the density of the displacement interpolant to be positive everywhere?*

For general Lagrangian actions, the answer to this question seems to be negative, but it is not clear that one can also construct counterexamples for, say, the basic quadratic Lagrangian. My personal guess would be that the answer is about the same as for the regularity theory: Positivity of the displacement interpolant is in general false except maybe for some particular manifolds satisfying an adequate structure condition.

Standard approximation procedures

In this last section I have gathered two useful approximation results which can be used in a great deal of problems where the probability measures are either noncompactly supported, or singular.

In Chapter 10 we have seen how to treat the Monge problem in noncompact situations, without any condition at infinity, thanks to the notion of approximate differentiability. However, in practice, to treat noncompact situations, the simplest solution is often to use again a truncation argument similar to the one used in the proof of approximate differentiability. The next proposition displays the standard scheme that one can use to deal with such situations.

Proposition 13.2 (Exhaustion of a noncompact transport by compact ones). *Let M be a Riemannian manifold, let $c = c(x, y)$ be a cost function coming from a Lagrangian $L(x, v, t)$ satisfying the classical assumptions of Definition 7.6;* **and strict convexity of L ??** *and let μ_0, μ_1 be two probability measures on M. Let π be an optimal transference plan*

between μ_0 and μ_1, let $(\mu_t)_{0\le t\le 1}$ be a displacement interpolation and let Π be a dynamical optimal transference plan such that $(e_0, e_1)_\# \Pi = \pi$, $(e_t)_\# \Pi = \mu_t$. Let Γ be the set of all action-minimizing curves, equipped with the topology of uniform convergence; and let $(K_\ell)_{\ell\in\mathbb{N}}$ be a sequence of compact sets in Γ, such that $\mathrm{Spt}(\Pi) \subset \cup K_\ell$. For ℓ large enough, $\Pi[K_\ell] > 0$; then define

$$Z_\ell := \Pi[K_\ell]; \qquad \Pi_\ell := \frac{1_{K_\ell}\Pi}{Z_\ell};$$

$$\mu_{t,\ell} := (e_t)_\# \Pi_\ell; \qquad \pi_\ell := (e_0, e_1)_\# \Pi_\ell;$$

and let c_ℓ be the restriction of c to $\mathrm{proj}(M \times M)(K_\ell)$. Then for each ℓ, $(\mu_{t,\ell})_{0\le t\le 1}$ is a displacement interpolation and π_ℓ is an associated optimal transference plan; $\mu_{t,\ell}$ is compactly supported, uniformly in $t \in [0,1]$; and the following monotone convergences hold true:

$$Z_\ell \uparrow 1; \qquad Z_\ell\, \pi_\ell \uparrow \pi; \qquad Z_\ell\, \mu_{t,\ell} \uparrow \mu_t; \qquad Z_\ell\, \Pi_\ell \uparrow \Pi.$$

If moreover μ_0 is absolutely continuous, then there exists a c-convex ψ such that π is concentrated on the graph of the transport $T : x \to (\nabla_x c)^{-1}(x, -\widetilde{\nabla}\psi(x))$, where the inverse is with respect to the second variable. Then for any ℓ, $\mu_{0,\ell}$ is also absolutely continuous, and the optimal transference plan π_ℓ is deterministic. Furthermore, there is a c_ℓ-convex function ψ_ℓ such that ψ_ℓ coincides with ψ everywhere on $C_\ell := \mathrm{proj}_M(\mathrm{Spt}(\pi_\ell))$. Futhermore, there is a set Z_ℓ such that $\mathrm{vol}\,[Z_\ell] = 0$ and for any $x \in C_\ell \setminus Z_\ell$, $\widetilde{\nabla}\psi(x) = \nabla\psi_\ell(x)$.

Still under the assumption that μ_0 is absolutely continuous, the measures $\mu_{t,\ell}$ are also absolutely continuous, and the optimal transport $T_{t_0\to t,\ell}$ between $\mu_{t_0,\ell}$ and $\mu_{t,\ell}$ is deterministic, for any given $t_0 \in [0,1)$ and $t \in [0,1]$. In addition, for any given $t_0 \in [0,1]$, one has

$$T_{t_0\to t,\ell} = T_{t_0\to t}, \qquad \mu_{t_0,\ell}\text{-almost surely},$$

where $T_{t_0\to t}$ is the optimal transport from μ_{t_0} to μ_t.

Proof of Proposition 13.2. The proof is quite similar to the argument used in the proof of uniqueness in Theorem 10.36 in a time-independent context. There is no problem to make this into a time-dependent version, since displacement interpolation behaves well under restriction, recall Theorem 7.27. The last part of the theorem follows from the fact that the map $T_{t_0\to t,\ell}$ can be written as $\gamma_{t_0} \to \gamma_t$. □

Now let us turn to the problem of approximating singular transport problems by smooth ones. If μ_0 and μ_1 are singular, there is a priori no uniqueness of the optimal transference plans, and actually there might be a large number (possibly uncountable) of them. However, the next theorem shows that singular optimal transference plans can always be approximated by nice ones.

Theorem 13.3 (Regularization of singular transport problems). *Let M be a Riemannian manifold, and $c : M\times M \to \mathbb{R}$ be a cost function induced by a Lagrangian $L(x,v,t)$ that is bounded below, C^2, and satisfies the classical assumptions of Definition 7.6, together with $\nabla_v^2 L > 0$. Let further μ_0 and μ_1 be two probability measures on M, such that the optimal transport cost between μ_0 and μ_1 is finite, and let π be an optimal transference plan between μ_0 and μ_1. Then there are sequences $(\mu_0^k)_{k\in\mathbb{N}}$, $(\mu_1^k)_{k\in\mathbb{N}}$ and $(\pi^k)_{k\in\mathbb{N}}$ such that*

(i) each π^k is an optimal transference plan between μ_0^k and μ_1^k, and any one of the probability measures μ_0^k, μ_1^k has a smooth, compactly supported density;

(ii) $\mu_0^k \to \mu_0$, $\mu_1^k \to \mu_1$, $\pi^k \to \pi$ in weak sense as $k \to \infty$.

Proof of Theorem 13.3. By Theorem 7.19, there exists a displacement interpolation $(\mu_t)_{0\le t\le 1}$ between μ_0 and μ_1; let γ be such that $\mu_t = \text{law}\,(\gamma_t)$. The assumptions on the cost function imply that action-minimizing curves solve a differential equation with Lipschitz coefficients, and therefore are uniquely determined by their initial position and velocity, a fortiori by their restriction to some time-interval $[0, t_0]$. So for any $t_0 \in (0, 1/2)$, by Theorem 7.27 (ii), $(\gamma_{t_0}, \gamma_{1-t_0})$ is the unique optimal coupling between μ_{t_0} and μ_{1-t_0}. Now it is easy to construct a sequence $(\mu_{t_0}^k)_{k\in\mathbb{N}}$ such that $\mu_{t_0}^k$ converges weakly to μ_{t_0} as $k \to \infty$, and each $\mu_{t_0}^k$ is compactly supported with a smooth density. (To construct such a sequence, first truncate to ensure the property of compact support, then localize to local charts by a partition of unity, and apply a regularization in each chart.) Similarly, construct a sequence $(\mu_{1-t_0}^k)_{k\in\mathbb{N}}$ such that $\mu_{1-t_0}^k$ converges weakly to μ_{1-t_0}, and each $\mu_{1-t_0}^k$ is compactly supported with a smooth density. Let $\pi_{t_0,1-t_0}^k$ be the unique optimal transference plan between μ_{t_0} and μ_{1-t_0}. By stability of optimal transport (Theorem 5.18), $\pi_{t_0,1-t_0}^k$ converges as $k \to \infty$ to $\pi_{t_0,1-t_0} = \text{law}\,(\gamma_{t_0}, \gamma_{1-t_0})$. Then by continuity of γ, the random variable $(\gamma_{t_0}, \gamma_{1-t_0})$ converges pointwise to (γ_0, γ_1) as $t_0 \to 0$; which implies that $\pi_{t_0,1-t_0}$ converges weakly to π. The conclusion follows by choosing $t_0 = 1/n$, $k = k(n)$ large enough. □

Equations of displacement interpolation

In Chapter 7, we understood that a curve $(\mu_t)_{0\le t\le 1}$ obtained by displacement interpolation could be seen as a solution of an action minimizing problem in the space of measures, and we wondered whether we could obtain some nice equations for these curves, and some nice action-minimizing principle in the space of curves. Here now is a possible answer. For simplicity I shall assume that there is enough control at infinity, that the notion of approximate differentiability can be dispended with.

Consider a displacement interpolation $(\mu_t)_{0\le t\le 1}$. As a consequence of Theorem 7.19, μ_t can be seen as the law of γ_t, where the random path $(\gamma_t)_{0\le t\le 1}$ satisfies the Euler-Lagrange equation (13.1), and so at time t has velocity $\xi_t(\gamma_t)$, where $\xi_t(x) := \big(\nabla_v L(x, \cdot, t)\big)^{-1}(\nabla\psi_t(x))$. By the formula of conservation of mass, recalled in the Appendix of the Introduction, μ_t satisfies

$$\frac{\partial \mu_t}{\partial t} + \nabla \cdot (\xi_t \mu_t) = 0$$

in the sense of distributions (be careful: w_t is not necessarily a gradient, unless L is quadratic). Then we can write down the **equations of displacement interpolation**:

$$\begin{cases} \dfrac{\partial \mu_t}{\partial t} + \nabla \cdot (\xi_t\, \mu_t) = 0 \\ \nabla_v L\big(x, \xi_t(x), t\big) = \nabla\psi_t(x); \\ \psi_0 \text{ is } c\text{-convex} \\ \partial_t \psi_t + L^*\big(x, \nabla\psi_t(x), t\big) = 0. \end{cases} \tag{13.5}$$

If the cost function is just the square of the distance, then these equations become

$$\begin{cases} \dfrac{\partial \mu_t}{\partial t} + \nabla \cdot (\xi_t \mu_t) = 0 \\ \xi_t(x) = \nabla\psi_t(x); \\ \psi_0 \text{ is } d^2/2\text{-convex} \\ \partial_t \psi_t + \dfrac{|\nabla\psi_t|^2}{2} = 0. \end{cases} \tag{13.6}$$

Finally, for the square of the Euclidean distance, this simplifies into

$$\begin{cases} \dfrac{\partial \mu_t}{\partial t} + \nabla \cdot (\xi_t \mu_t) = 0 \\ \xi_t(x) = \nabla \psi_t(x); \\ \Psi(x) := x + \psi_0(x) \text{ is a convex function of } x \\ \partial_t \psi_t + \dfrac{|\nabla \psi_t|^2}{2} = 0. \end{cases} \tag{13.7}$$

Quadratic cost function

In a context of Riemannian geometry, it is natural to focus on the quadratic Lagrangian cost function, or equivalently on the cost function $c(x,y) = d(x,y)^2$, and consider the Wasserstein space $P_2(M)$. This will be the core of all the transport proofs in Part II of these notes, and so a key role will be played by $d^2/2$-convex functions. In Part III we shall consider metric structures than are not Riemannian, but still the square of the distance will be the only cost function. So in the remaining of this chapter I shall focus on that particular cost.

The class of $d^2/2$-convex functions might look a bit mysterious, and if they are so important it would be good to have simple characterizations of them. If ψ is $d^2/2$-convex, then $z \to \psi(z) + d(z,y)^2/2$ should be minimum at x when $y = \exp_x(\nabla \psi(x))$. If in addition ψ is twice differentiable at x, then necessarily

$$\nabla^2 \psi(x) \geq \nabla^2 \left[\frac{d(\cdot, \exp_x \nabla \psi(x))^2}{2} \right](x). \tag{13.8}$$

However, this is only a necessary condition, and I don't know if it implies $d^2/2$-convexity, alone or together with some other reasonably simple condition.

On the other hand, there is a simple and useful criterion according to which sufficiently small functions are $d^2/2$-convex. This statement will guarantee in particular that any tangent vector $v \in TM$ can be represented as the gradient of a $d^2/2$-convex function.

Theorem 13.4 (C^2-small functions are $d^2/2$-convex). *Let M be a Riemannian manifold, and let K be a compact subset of M. Then, there is $\varepsilon > 0$ such that any function $\psi \in C^2_c(M)$ satisfying*

$$\mathrm{Spt}(\psi) \subset K, \qquad \|\psi\|_{C^2_b} \leq \varepsilon$$

is $d^2/2$-convex.

Example 13.5. If $M = \mathbb{R}^n$, then ψ is $d^2/2$-convex as soon as $\nabla^2 \psi \geq -I_n$.

Proof. Let (M, g) be a Riemannian manifold, and let K be a compact subset of M. Let $K' = \{x \in M;\ d(x,K) \leq 1\}$. For any $y \in M$, the Hessian of $x \to d(x,y)^2/2$ is equal to I_n (or, more rigorously, to the metric tensor g) at $x = y$; so by compactness one may find $\delta > 0$ such that the Hessian of $x \to d(x,y)^2/2$ remains larger than $I_n/2$ as long as y stays in K' and $d(x,y) < 2\delta$. Without loss of generality, $\delta < 1/2$.

Now let ψ be supported in K, and such that

$$\forall x \in M \qquad |\psi(x)| < \frac{\delta^2}{2}, \quad |\nabla^2 \psi(x)| < \frac{1}{2};$$

write

$$f_y(x) = \psi(x) + \frac{d(x,y)^2}{2}.$$

If $y \in K'$ and $d(x,y) \geq \delta$, then obviously $f_y(x) > f_y(y)$; so the minimum of f_y can be achieved only in $B_\delta(y)$. If there are two distinct such minima, say x_0 and x_1, then we can join them by a geodesic $(\gamma_t)_{0\leq t\leq 1}$ which stays within $B_{2\delta}(y)$ and then the function $t \to f_y(\gamma_t)$ is uniformly convex (because f_y is uniformly convex in $B_{2\delta}(y)$, and minimum at $t = 0$ and $t = 1$, which is impossible.

If on the other hand $y \notin K'$, then the minimum of f_y can be achieved only if $d(x,y) < 1$, which imposes $\psi(x) = 0$, so the minimum is necessarily at $y = x$.

The conclusion is that in any case, f_y has exactly one minimum, which lies in $B_\delta(y)$. We shall denote it by $x = T(y)$, and it is characterized as the unique solution of the equation

$$\nabla\psi(x) + \nabla_x \frac{d(x,y)^2}{2} = 0, \tag{13.9}$$

where x is the unknown.

Let x be arbitrary in M, and $y = \exp_x(\nabla\psi(x))$. Then (as a consequence of the first variation formula), $\nabla_x[d(x,y)^2/2] = -\nabla\psi(x)$, so equation (13.9) hold true, and $x = T(y)$. This means that, with the notation $c(x,y) = d(x,y)^2/2$, one has $\psi^c(y) = \psi(x) + c(x,y)$. Then $\psi^{cc}(x) = \sup[\psi^c(y) - c(x,y)] \geq \psi(x)$. Since x is arbitrary, actually we have shown that $\psi^{cc} \geq \psi$; but the converse inequality is always true, so $\psi^{cc} = \psi$, and then ψ is c-convex. □

Remark 13.6. The end of the proof took advantage of a general principle, independent of the particular cost c: If there is a *surjective* map T such that $f_y : x \to \psi(x) + c(x,y)$ is minimum at $T(y)$, then ψ is c-convex.

The structure of $P_2(M)$

It was one of the striking discoveries of the end of the nineties that the differentiable structure on a Riemannian manifold M induces a kind of differentiable structure in the space $P_2(M)$. This idea takes substance from the following remarks: All of the path $(\mu_t)_{0\leq t\leq 1}$ is determined from the initial velocity field $\xi_0(x)$, which in turn is determined by $\nabla\psi$ as in (13.4). So it is natural to think of the function $\nabla\psi$ as a kind of "initial velocity" for the path (μ_t). The conceptual shift here is about the same as when we decided that μ_t could be seen either as the law of a random minimizing curve at time t, or as a path in the space of measures: Now we decide that $\nabla\psi$ can be seen either as the field of the initial velocities of our minimizing curves, or as the (abstract) velocity of the path μ_t at time $t = 0$.

There is an abstract notion of tangent space $T_x\mathcal{X}$ (at point x) to a metric space $(\mathcal{X}, d)$: in technical language, this is the *pointed Gromov-Hausdorff limit* of the rescaled space. This actually is a rather natural notion: fix your point x, and zoom on it, by multiplying all distances by a large factor ε^{-1}, while keeping x fixed. This gives a new metric space $\mathcal{X}_{x,\varepsilon}$, and if one is not too curious about what happens far away from x, then the space $\mathcal{X}_{x,\varepsilon}$ might converge in some nice sense to some limit space, that may not be a vector space, but in any case is a cone. If that limit space exists, it is said to be the tangent space (or tangent cone) to $\mathcal{X}$ at x.

In terms of that construction, the intuition sketched above is indeed correct: let $P_2(M)$ be the metric space consisting of probability measures on M, equipped with the Wasserstein distance W_2. If μ is absolutely continuous, then the tangent cone $T_\mu P_2(M)$ exists

and can be identified isometrically with the closed vector space generated by $d^2/2$-convex functions ψ, equipped with the norm

$$\|\nabla\psi\|_{L^2(\mu;TM)} := \left(\int_M |\nabla\psi|_x^2 \, d\mu(x)\right)^{1/2}.$$

(Actually, in view of Theorem 13.4, this is the same as the vector space generated by all smooth, compactly supported gradients with respect to that norm.) With what we know about optimal transport, this theorem is not that hard to prove, but this would require a bit too much of geometric machinery for now. Instead, I shall spend some time on an important result by Ambrosio, Gigli and Savaré, showing that any Lipschitz curve in the space $P_2(M)$ admits a *velocity* (which for all t lives in the tangent space at μ_t). Surprisingly, the proof will not require absolute continuity. I state the theorem on a compact Riemannian manifold, but the exact same proof would work in $\mathbb{R}^n$. For a general Riemannian manifold, it might be that some conditions at infinity are needed.

Theorem 13.7 (Representation of Lipschitz measure-valued curves). *Let M be a smooth complete Riemannian manifold, and let $P_2(M)$ be the metric space of all probability measures on M, with a finite second moment, equipped with the metric W_2. Let further $(\mu_t)_{0\le t\le 1}$ be a Lipschitz-continuous path in $P_2(M)$:*

$$W_2(\mu_s, \mu_t) \le L\,|t-s|.$$

For all $t \in [0,1]$, let H_t be the Hilbert space generated by gradients of continuously differentiable, compactly supported ψ:

$$H_t := \overline{\mathrm{Vect}\big(\{\nabla\psi;\quad \psi \in C_c^1(M)\}\big)}^{L^2(\mu_t;TM)}.$$

Then there exists a measurable vector field $\xi_t(x) \in L^\infty(dt;\, L^2(d\mu_t(x)))$, $\mu_t(dx)\,dt$-almost everywhere unique, such that $\xi_t \in H_t$ for all t (i.e. the velocity field really is tangent along the path), and

$$\partial_t \mu_t + \nabla\cdot(\xi_t \mu_t) = 0 \tag{13.10}$$

in weak sense.

Conversely, if the path $(\mu_t)_{0\le t\le 1}$ satisfies (13.10) *for some measurable vector field $(\xi_t(x))$ whose $L^2(\mu_t)$-norm is bounded almost surely in t, then (μ_t) is a Lipschitz-continuous curve.*

The proof of Theorem 13.7 requires a bit of analytical tools, and the reader might skip it at first reading.

Proof of Theorem 13.7. Let $\psi : M \to \mathbb{R}$ be C^1 function, with Lipschitz constant at most 1. For all $s < t$ in $[0,1]$,

$$\left|\int_M \psi\, d\mu_t - \int_M \psi\, d\mu_s\right| \le W_1(\mu_s,\mu_t) \le W_2(\mu_s,\mu_t). \tag{13.11}$$

In particular, $\zeta(t) := \int_M \psi\, d\mu_t$ is a Lipschitz function of t. By a classical theorem of real analysis, the time-derivative of ζ exists for almost all time $t \in [0,1]$. Let then $\pi_{s,t}$ be an optimal transference plan between μ_s and μ_t (for the squared distance cost function). Let

$$\Psi(x,y) := \begin{cases} \dfrac{|\psi(x)-\psi(y)|}{d(x,y)} & \text{if } x \ne y \\[2ex] |\nabla\psi(x)| & \text{if } x = y. \end{cases}$$

Obviously Ψ is bounded by 1, and moreover it is upper semi-continuous.

Then, if t is a differentiability point of ζ,

$$\begin{aligned}
\left|\frac{d}{dt}\int \psi\, d\mu_t\right| &\le \liminf_{\varepsilon\downarrow 0} \frac{1}{\varepsilon}\left|\int \psi\, d\mu_t - \int \psi\, d\mu_{t+\varepsilon}\right| \\
&\le \liminf_{\varepsilon\downarrow 0} \frac{1}{\varepsilon}\int |\psi(y)-\psi(x)|\, d\pi_{t,t+\varepsilon}(x,y) \\
&\le \liminf_{\varepsilon\downarrow 0} \sqrt{\int \Psi(x,y)^2\, d\pi_{t,t+\varepsilon}(x,y)}\ \frac{\sqrt{\int d(x,y)^2\, d\pi_{t,t+\varepsilon}(x,y)}}{\varepsilon} \\
&= \liminf_{\varepsilon\downarrow 0} \sqrt{\int \Psi(x,y)^2\, d\pi_{t,t+\varepsilon}(x,y)}\ \frac{W_2(\mu_t,\mu_{t+\varepsilon})}{\varepsilon} \\
&\le \liminf_{\varepsilon\downarrow 0} \sqrt{\int \Psi(x,y)^2\, d\pi_{t,t+\varepsilon}(x,y)}\ L.
\end{aligned}$$

Since Ψ is upper semi-continuous and $\pi_{t,t+\varepsilon}$ converges weakly to $\delta_{x=y}$ (the trivial transport plan where nothing moves) as $\varepsilon \downarrow 0$, it follows that

$$\begin{aligned}
\left|\frac{d}{dt}\int \psi\, d\mu_t\right| &\le L\sqrt{\int |\Psi(x,x)|^2\, d\mu_t(x)} \\
&= L\sqrt{\int |\nabla\psi(x)|^2\, d\mu_t(x)}.
\end{aligned}$$

Now the key thing is that $(d/dt)\int(\psi + C)\, d\mu_t$ does not depend on the constant C. This shows that $(d/dt)\int \psi\, d\mu_t$ really is a functional of $\nabla\psi$, and obviously linear functional. Then the above estimate shows that this functional is *continuous* with respect to the norm in $L^2(d\mu_t)$.

Actually, this is not completely rigorous, since this functional is only defined for almost all t, and "almost all" here might depend on ψ. Here is a way to make things rigorous: Let $\mathcal{L}$ be the set of all Lipschitz functions ψ on M with Lipschitz constant at most 1, such that, say, $\psi(x_0) = 0$, where $x_0 \in M$ is arbitrary but fixed once for all, and ψ is supported in a fixed compact $K \subset M$. The set $\mathcal{L}$ is compact in the norm of uniform convergence, and admits a dense sequence $(\psi_k)_{k\in\mathbb{N}}$. By a regularization argument, one can assume that all those functions are actually of class C^1. For each ψ_k, we know that $(d/dt)\int \psi_k\, d\mu_t$ is differentiable for almost all $t \in [0,1]$; and since there are only countably many ζ_k's, we know that for almost every t, each ζ_k is differentiable at time t. Now the map $(d/dt)\int v\, d\mu_t$ is well-defined at each of these times t, for all ξ in the vector space H_t generated by all the ψ_k's; and it is continuous if that vector space is equipped with the $L^2(d\mu_t)$ norm. It follows from the Riesz representation theorem that for each differentiability time t there exists a unique vector $\xi_t \in H_t \subset L^2(d\mu_t)$, with norm at most L, such that

$$\frac{d}{dt}\int \psi\, d\mu_t = \int \xi_t \cdot \nabla\psi\, d\mu_t. \tag{13.12}$$

This identity should hold true for any ψ_k, and by density it should also hold true for any $\psi \in C^1(M)$, supported in K.

Let $C^1_K(M)$ be the set of $\psi \in C^1(M)$ that are supported in K. We just showed that there is a negligible set of times, τ_K, such that (13.12) holds true for all $\psi \in C^1_K(M)$ and $t \notin \tau_K$.

Now choose an increasing family of compact sets $(K_m)_{m\in\mathbb{N}}$, with $\cup K_m = M$, so that any compact set is included in some K_m. Then (13.12) will hold true for all $\psi \in C^1_c(M)$, as soon as t does not belong to the union of τ_{K_m}, which is still a negligible set of times.

But equation (13.12) really is the weak formulation of (13.10). Since v_t is uniquely determined in $L^2(d\mu_t)$, for almost all t, actually the vector field $v_t(x)$ is $d\mu_t(x)\,dt$-uniquely determined.

To conclude the proof of the theorem, it only remains to prove the converse implication. Let (μ_t) and (ξ_t) solve (13.10). By the equation of conservation of mass, $\mu_t = \text{law}\,(\gamma_t)$, where γ_t is a (random) solution of

$$\dot\gamma_t = \xi_t(\gamma_t).$$

Let $s < t$ be any two times in $[0,1]$. From the formula

$$d(\gamma_s,\gamma_t)^2 = (t-s)\inf\Big\{\int_s^t |\dot\zeta_\tau|^2\,d\tau;\quad \zeta_s=\gamma_s,\ \zeta_t=\gamma_t\Big\},$$

we deduce

$$d(\gamma_s,\gamma_t)^2 \le (t-s)\int_s^t |\dot\gamma_\tau|^2\,d\tau \le (t-s)\int_s^t |\xi_t(\gamma_t)|^2\,d\tau.$$

So

$$\mathbb{E}\,d(\gamma_s,\gamma_t)^2 \le (t-s)\int_s^t |\xi_\tau(x)|^2\,d\mu_\tau(x)\,d\tau \le (t-s)^2\|\xi\|_{L^\infty(dt;\,L^2(d\mu_t))}.$$

In particular

$$W_2(\mu_s,\mu_t)^2 \le \mathbb{E}\,d(\gamma_s,\gamma_t)^2 \le L^2(t-s)^2,$$

where L is an upper bound for the norm of ξ in $L^\infty(L^2)$. This concludes the proof of Theorem 13.7. □

Remark 13.8. With hardly more work, the preceding theorem can be extended to cover paths that are absolutely continuous of order 2, in the sense defined on p. 7. Then of course the velocity field will not live in $L^\infty(dt;\,L^2(d\mu_t))$, but in $L^2(d\mu_t\,dt)$.

Observe that in a displacement interpolation, the initial measure μ_0 and the initial velocity field $\nabla\psi_0$ uniquely determine the final measure μ_1: this implies that *geodesics in $P_2(M)$ are nonbranching*, in the strong sense that their initial position and velocity determine uniquely their final position.

Finally, we can now derive an "explicit" formula for the action functional determining displacement interpolations as minimizing curves. Let $\mu = (\mu_t)_{0\le t}$ be any Lipschitz (or absolutely continuous) path in $P_2(M)$; let $\xi_t(x) = \nabla\psi_t(x)$ be the associated time-dependent velocity field. By the formula of conservation of mass, μ_t can be interpreted as the law of γ_t, where γ is a random solution of $\dot\gamma_t = \xi_t(\gamma_t)$. Now define

$$\mathbb{A}(\mu) := \inf \int_0^1 \mathbb{E}_{\,\mu_t} |\xi_t(\gamma_t)|^2\,dt \tag{13.13}$$

where the infimum is taken over all possible realizations of the random curves γ. By Fubini's theorem,

$$\begin{aligned}\mathbb{A}(\mu) = \inf \mathbb{E}\int_0^1 |\xi_t(\gamma_t)|^2\,dt &= \inf \mathbb{E}\int_0^1 |\dot\gamma_t|^2\,dt \\ &\ge \mathbb{E}\inf\int_0^1 |\dot\gamma_t|^2\,dt \qquad\qquad = \mathbb{E}\,d(\gamma_0,\gamma_1)^2,\end{aligned}$$

and the infimum is achieved if only if the coupling (γ_0, γ_1) is minimal, and the curves γ are (almost surely) action-minimizing. This shows that displacement interpolations are characterized as the minimizing curves for the action $\mathbb{A}$. Actually $\mathbb{A}$ is the same as the action appearing in Theorem 7.19 (iii), the only improvement is that now we have given it a more explicit form in terms of vector fields.

The expression (13.13) can be made slightly more explicit by noting that the optimal choice of velocity field is the one provided by Theorem 13.7, which is gradient, so we may restrict the action functional to gradient velocity fields:

$$\mathbb{A}(\mu) := \int_0^1 \mathbb{E}_{\,\mu_t} |\nabla \psi_t|^2 \, dt; \qquad \frac{\partial \mu_t}{\partial t} + \nabla \cdot (\nabla \psi_t \, \mu_t) = 0. \tag{13.14}$$

Note the formal resemblance with a Riemannian structure: what the formula above says is

$$W_2(\mu_0, \mu_1)^2 = \inf \int_0^1 \|\dot{\mu}_t\|^2_{T_{\mu_t} P_2} \, dt, \tag{13.15}$$

where the norm on the tangent space $T_\mu P_2$ is defined by

$$\begin{aligned} \|\dot{\mu}\|^2_{T_\mu P_2} &= \inf \left\{ \int |v|^2 \, d\mu; \qquad \dot{\mu} + \nabla \cdot (v\mu) = 0 \right\} \\ &= \int |\nabla \psi|^2 \, d\mu; \qquad \dot{\mu} + \nabla \cdot (\nabla \psi \, \mu) = 0. \end{aligned}$$

There is an appealing physical interpretation, which really is an infinitesimal version of the optimal transport problem. Imagine that you observe the (infinitesimal) evolution of the density of particles moving in a continuum, but don't know the actual velocities of these particles. There might be many velocity fields that are compatible with the observed evolution of density (many solutions of the continuity equation). Among all the possible solutions, select the one with minimum kinetic energy. This energy is (up to a factor 2) the square norm of your infinitesimal evolution.

Bibliographical Notes

Formula (13.8) appears in [118]. It has an interesting consequence which can be described as follows: On a Riemannian manifold, the optimal transport starting from an absolutely continuous probability measure *almost never hits the cut locus*; that is, the set of x such that the image $T(x)$ belongs to the cut locus of x is of zero probability. Although we already know (by uniqueness of the displacement interpolation, for instance) that almost surely, x and $T(x)$ are joined by a unique geodesic, this alone does not imply that the cut locus is almost never hit, because it is possible that y belongs to the cut locus of x and still x and y are joined by a unique minimizing geodesic. (Recall the discussion after Problem 8.9.) But Cordero-Erausquin, McCann and Schmuckenschläger show that if such is the case, then $d(x,z)^2/2$ fails to be semi-convex at $z = y$. On the other hand, it follows from Alexandrov's second differentiability theorem (recalled in an Appendix to Chapter 14) that ψ is twice differentiable almost everywhere, and then formula (13.8), suitably interpreted, says that $d(x,\cdot)^2/2$ is semi-convex at $T(x)$ whenever ψ is twice differentiable at x. I did not include the proof of this result in these notes, because it uses more advanced Riemannian geometry technology.

At least in the Euclidean case, the explicit formulas for geodesic curves and action in the space of measures were known to Brenier, no later than the mid-nineties, but from a

formal point of view. Otto [290] took a conceptual step forward by considering formally $P_2(M)$ as an infinite-dimensional Riemannian manifold, in view of formula (13.15). For some time it was used as a purely formal, yet quite useful, heuristics (as in [292], or later in this course). It is only recently that rigorous constructions were performed in several research papers, including [102, 15, 98, 247]. The treatment developed in this chapter relies heavily on the work of Ambrosio, Gigli and Savaré [15] (in $\mathbb{R}^n$). The most geometric study is probably the one in [247, Appendix A]; but see also [15, Section 12.4].

In his PhD, Agueh [1] studied what happens to this picture when the quadratic cost function is replaced by a power law cost function $|x - y|^p$ $(p > 1)$. Bernard and Buffoni [47] made a partial investigation of the case of smooth, strictly convex Lagrangian on a manifold, but not so much from the point which was developed in this end of this chapter.

Displacement interpolation in presence of boundaries becomes quite tricky. See Otto [290] for some partial study in a bounded open set of $\mathbb{R}^n$ with C^2 boundary.

So far, the great majority of applications of optimal transport to problems of applied mathematics have taken place in Euclidean setting, but more recently some "genuinely Riemannian" applications have started to pop out. There was a quite original suggestion to use optimal transport in a three-dimensional Riemannian manifold (actually, a cube equipped with a varying metric) related to image perception and the matching of pictures with different contrasts [134]. In a meteorological context, it is natural to consider the sphere (as a model of the Earth), and in the study of the semi-geostrophic system one is naturally led to optimal transport on the sphere [124, 122]; actually, it is even natural to consider a conformal change of metric which "pinches" the sphere along its equator [124]! For completely different reasons, optimal transport on the sphere was recently used by Otto and Tzavaras [291] in the study of a coupled fluid-polymer model.

Part II

Optimal transport and Riemannian geometry

This second part is devoted to the influence of geometry on the qualitative properties of optimal transport. It will be shown that the geometry of the manifold influences the qualitative properties of optimal transport, and that this can be quantified in particular by the influence of *Ricci curvature bounds* on *displacement convexity*, that is the convexity properties of certain well-chosen functionals along displacement interpolation.

Chapter 14 is a preliminary chapter devoted to a short exposition about the main properties of Ricci curvature. It is sufficiently self-contained that the reader should understand all the rest without having to consult any extra source on Riemannian geometry. The estimates in this chapter will be used only in Chapters 15, 16 and 17.

Chapter 15 is devoted to a powerful formal differential calculus on the Wasserstein space, found by Otto.

Chapters 16 and 17 establish the main relations between displacement convexity and Ricci curvature. Not only do Ricci curvature bounds imply certain properties of displacement convexity, but conversely those properties in fact characterize Ricci curvature bounds. The results in this chapter will play a key role in the rest of the course.

In Chapters 18 to 22 the main theme will be that many classical properties of Riemannian manifolds, that come from Ricci curvature estimates, can be conveniently derived from displacement convexity techniques. This includes in particular estimates about the growth of the volume of balls, Sobolev-type inequalities, concentration inequalities, and Poincaré inequalities.

Then in Chapter 23 it is explained how one can define certain *gradient flows* in the Wasserstein space, and recover in this way certain well-known equations such as the heat equation. In Chapter 24, some of the functional inequalities that were established in the previous chapters are then applied to the study of these equations; and conversely, gradient flows provide alternative proofs to some of these inequalities, as shown in Chapter 25.

14

Ricci curvature

Curvature is a generic name to designate a local invariant of a metric space that quantifies the deviation of this space from being Euclidean. (Here "local invariant" means a quantity which is invariant under local isometries.) It is standard to define and study curvature mainly on Riemannian manifolds, for in that setting definitions are rather simple, and the Riemannian structure allows for "explicit" computations. In all this chapter, M will stand for a smooth Riemannian manifold, equipped with a smooth metric g. Unless explicitly mentioned, M will also be assumed to be complete.

The most popular curvatures are: the **sectional curvature** σ (for each point x and each plane $P \subset T_xM$, $\sigma_x(P)$ is a number), the **Ricci curvature** Ric (for each point x, Ric_x is a quadratic form on the tangent space T_xM), and the **scalar curvature** S (for each point x, S_x is a number). All of them can be obtained by reduction of the **Riemann curvature tensor**. The latter is easy to define: If ∇_X stands for the covariant derivation along the vector field X, then

$$\mathrm{Riem}(X,Y) := \nabla_Y \nabla_X - \nabla_X \nabla_Y + \nabla_{[X,Y]};$$

but it is notoriously difficult to get some intuition about its meaning, even for specialists. The Riemann curvature can be thought of as a tensor with four indices; it can be expressed in coordinates as a nonlinear function of the Christoffel symbols and their partial derivatives.

Of these three notions of curvature (sectional, Ricci, scalar), the sectional one is the most precise; in fact the knowledge of all sectional curvatures is equivalent to the knowledge of the Riemann curvature. Then the Ricci curvature is obtained by "tracing" the sectional curvature: If e is a given unit vector in T_xM and $(e, e_2, \ldots, e_n)$ is an orthonormal basis of T_xM, then $\mathrm{Ric}_x(e,e) = \sum \sigma_x(P_j)$, where P_j $(j = 2, \ldots, n)$ is the plane generated by $\{e, e_j\}$. Finally, the sectional curvature is the trace of the Ricci curvature. So a control on the sectional curvature is stronger than a control on the Ricci curvature, which in turn is stronger than a control on the scalar curvature.

For a surface (manifold of dimension 2), these three notions reduce to just one, which is the Gauss curvature and whose definition is elementary. Let us first describe it from an extrinsic point of view. Let M be a two-dimensional submanifold of $\mathbb{R}^3$. In the neighborhood of a point x, choose a unit normal vector $n = n(y)$, then this defines locally a smooth map n with values in $S^2 \subset \mathbb{R}^3$. The tangent spaces T_xM and $T_{n(x)}S^2$ are parallel planes in $\mathbb{R}^3$, which can be identified unambiguously. So the determinant of the differential of n can also be defined without ambiguity, and this determinant is called the curvature. The fact that this quantity is invariant under isometries is one of Gauss's most famous results, a *tour de force* at the time. (To appreciate this theorem, the reader might try to prove it by elementary means.)

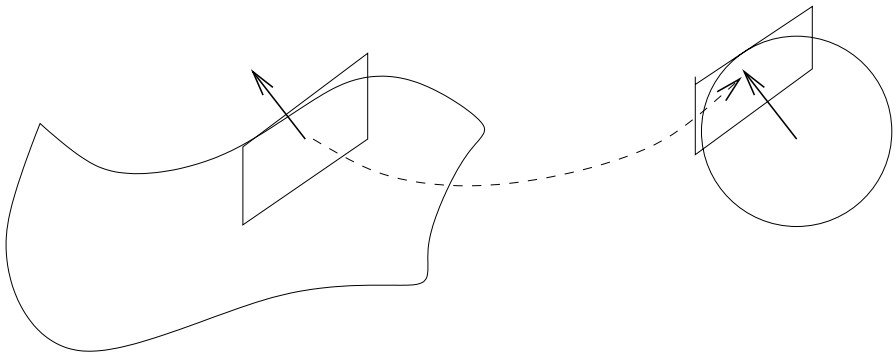


Fig. 14.1. The dashed line gives the recipe for the construction of the Gauss map; its Jacobian determinant is the Gauss curvature.

As an illustration of this theorem: If you hold a sheet of paper straight, then its equation (as an embedded surface in $\mathbb{R}^3$, and assuming that it is infinite) is just the equation of a plane, so obviously it is not curved. Fine, but now bend the sheet of paper so that it looks like valleys and mountains, write down the horrible resulting equations, give it to a friend and ask him whether it is curved or not. One thing he can do is compute the Gauss curvature from your horrible equations, find that it is identically 0, and deduce that your surface was not curved at all. Well, it looked curved as a surface which was embedded in $\mathbb{R}^3$, but from an *intrinsic* point of view it was not: A tiny creature living on the surface of the sheet, unable to measure the lengths of curves going outside of the surface, would never have noticed that you bent the sheet.

To construct isometries from (M, g) to something else, pick up any diffeomorphism $\varphi : M \to M'$, and equip $M' = \varphi(M)$ with the metric $g' = (\varphi^{-1})^* g$, defined by $g'_x(v) = g_{\varphi^{-1}(x)}(d_x\varphi^{-1}(v))$. Then φ is an isometry between (M, g) and (M', g'). Gauss' theorem says that the curvature computed in (M, g) and the curvature computed in (M', g') are the same, modulo obvious changes (the curvature at point x along a plane P should be compared with the curvature at $\varphi(x)$ along a plane $d_x\varphi(P)$). This is why one often says that the curvature is "invariant under the action of diffeomorphisms".

Curvature is intimately related to the local behavior of geodesics. The general rule is that, in presence of positive curvature, geodesics have a tendency to *converge* (at least in short time), while in presence of negative curvature they have a tendency to *diverge*. This tendency can usually be felt only at second or third order in time: at first order, the convergence or divergence of geodesics is dictated by the initial conditions. So if, on a space of (strictly) positive curvature, you start two geodesics from the same point with velocities pointing in different directions, the geodesics will start to diverge, but then the tendency to diverge will diminish. Here is a more precise statement, which will show at the same time that the Gauss curvature is an intrinsic notion: From a point $x \in M$, start two constant-speed geodesics with unit speed, and respective velocities v, w. The two curves will then spread apart; let $\delta(t)$ be the distance between their respective positions at time t. In a first approximation, $\delta(t) \simeq \sqrt{2(1 - \cos\theta)}\, t$, where θ is the angle between v and w (this is the same formula as in Euclidean space). But a more precise study shows that

$$\delta(t) = \sqrt{2(1 - \cos\theta)}\, t \left(1 - \frac{\kappa_x \cos^2(\theta/2)}{6}\, t^2 + O(t^4)\right), \tag{14.1}$$

where κ_x is the Gauss curvature at x.

Once the intrinsic nature of the Gauss curvature has been established, it is easy to define the notion of sectional curvature for Riemannian manifolds of any dimension, embedded

or not: If $x \in M$ and $P \subset T_xM$, define $\sigma_x(P)$ as the Gauss curvature of the surface which is obtained as the image of P by the exponential map $\exp_x$ (that is, the collection of all geodesics starting from x with a velocity in P). Another equivalent definition is by reduction of the Riemann curvature tensor: If $\{u, v\}$ is an orthonormal basis of P, then $\sigma_x(P) = \langle \mathrm{Riem}\,(u, v) \cdot u,\, v \rangle$.

It is obvious from the first definition that the unit two-dimensional sphere S^2 has curvature $+1$, and that the Euclidean plane $\mathbb{R}^2$ has curvature 0. More generally, the sphere $S^n(R)$, with dimension n and radius R, has constant sectional curvature $1/R^2$, while the n-dimensional Euclidean space $\mathbb{R}^n$ has curvature 0. The other classical example is the hyperbolic plane, say $\mathbb{H}^n(R) = \{(x, y) \in \mathbb{R}^{n-1} \times (0, +\infty)\}$ equipped with the metric $R^2(dx^2 + dy^2)/y^2$, which has constant sectional curvature $-1/R^2$. These three families (spheres, Euclidean, hyperbolic) constitute the only connected Riemannian manifolds with constant sectional curvature, and they play an important role as comparison spaces.

The qualitative properties of optimal transport are also (of course) related to the behavior of geodesics, and so it is natural to believe that curvature has a strong influence on the solution of optimal transport. Conversely, some curvature properties can be read off on the solution of optimal transport. At the time of writing, these links have been best understood in terms of *Ricci curvature*; so this is the point of view that will be developed in the sequel.

This chapter is a tentative crash course on Ricci curvature. Hopefully, a reader who has never heard about that topic before should, by the end of the chapter, know enough about it to understand all the rest of the notes. This is by no means a complete course, since most proofs will only be sketched and many basic results will be taken for granted.

In practice, Ricci curvature usually appears from two points of view: (a) estimates of the *Jacobian determinant of the exponential map*; (b) *Bochner's formula*. These are two complementary points of view on the same phenomenon, and it is useful to know both. Before going on, I shall make some preliminary remarks about Riemannian calculus at second order, for functions which are not necessarily smooth.

Preliminary: second-order differentiation

All curvature calculations involve second-order differentiation of certain expressions. The notion of covariant derivation lends itself well to those computations. A first thing to know is that the exchange of derivatives is still possible. To express this properly, consider a parametrized surface $(s, t) \to \gamma(s, t)$ in M, and write d/dt (resp. d/ds) for the differentiation along γ, viewed as a function of t with s frozen (resp. as a function of s with t frozen); and D/Dt (resp. D/Ds) for the corresponding covariant differentiation. Then, if $F \in C^2(M)$, one has

$$\frac{D}{Ds}\left(\frac{dF}{dt}\right) = \frac{D}{Dt}\left(\frac{dF}{ds}\right). \tag{14.2}$$

Also a crucial concept is that of **Hessian operator**. If f is twice differentiable on $\mathbb{R}^n$, its Hessian matrix is just $(\partial^2 f/\partial x_i\, \partial x_j)_{1 \le i,j \le n}$ of all second-order partial derivatives. Now if f is defined on a Riemannian manifold M, the Hessian operator at x is the linear operator $\nabla^2 f(x) : T_xM \to T_xM$ defined by the identity

$$\nabla^2 f \cdot v = \nabla_v(\nabla f).$$

(Recall that ∇_v stands for the covariant derivation in the direction v.) In short, $\nabla^2 f$ is the covariant gradient of the gradient of f.

A convenient way to compute the Hessian of a function is to differentiate it twice along a geodesic path. Indeed, if $(\gamma_t)_{0\le t\le 1}$ is a geodesic path, then

$$\frac{d^2}{dt^2}\, f(\gamma_t) = \frac{d}{dt}\langle \nabla f(\gamma_t),\, \dot\gamma_t\rangle = \Big\langle \nabla_{\dot\gamma}\nabla f(\gamma_t),\, \dot\gamma_t\Big\rangle + \Big\langle \nabla f(\gamma_t),\, \nabla_{\dot\gamma}\dot\gamma_t\Big\rangle \quad = \langle \nabla^2 f(\gamma_t)\cdot\dot\gamma_t,\, \dot\gamma_t\rangle.$$

In other words, if $\gamma_0 = x$ and $\dot\gamma_0 = v \in T_xM$, then

$$f(\gamma_t) = f(x) + t\langle \nabla f(x),\, v\rangle + \frac{t^2}{2}\,\langle \nabla^2 f(x)\cdot v,\, v\rangle + o(t^2). \tag{14.3}$$

This identity can actually be used to define the Hessian operator.

A similar computation shows that for any two tangent vectors u, v at x,

$$\frac{D}{Ds}\left(\frac{d}{dt}\, f\big(\exp_x(su+tv)\big)\right) = \langle \nabla^2 f(x)\cdot u,\, v\rangle, \tag{14.4}$$

where $\exp_x v$ is the value at time 1 of the constant speed geodesic starting from x with velocity v. Identity (14.4) together with (14.2) shows that if $f \in C^2(M)$, then $\nabla^2 f(x)$ is a symmetric operator, that is $\langle \nabla^2 f(x)\cdot u,\, v\rangle_x = \langle \nabla^2 f(x)\cdot v,\, u\rangle_x$. In that case it will be often convenient to think of $\nabla^f(x)$ as a quadratic form on T_xM.

The Hessian operator is related to another fundamental second-order differential operator, the **Laplacian**, or Laplace–Beltrami operator. The Laplacian can be defined as the trace of the Hessian:

$$\Delta f(x) = \mathrm{tr}\,(\nabla^2 f(x)).$$

Another possible definition for the Laplacian is

$$\Delta f = \nabla\cdot(\nabla f),$$

where $\nabla\cdot$ is the **divergence** operator, defined as the negative of the adjoint of the gradient in $L^2(M)$: More explicitly, if ξ is a C^1 vector field on M, then its divergence is defined by

$$\forall \zeta \in C_c^\infty(M), \qquad \int_M (\nabla\cdot\xi)\,\zeta\, d\mathrm{vol} \,= -\int_M \xi\cdot\nabla\zeta\, d\mathrm{vol}\,.$$

Both definitions are equivalent; in fact, more generally the divergence of a vector field ξ coincides with the trace of the covariant gradient of ξ. When $M = \mathbb{R}^n$, Δf is given by the usual expression $\sum \partial^2_{ii} f$. More generally, in coordinates, the Laplacian reads

$$\Delta f = (\det g)^{-1/2} \sum_i \partial_i\big((\det g)^{1/2} g^{ij}\,\partial_j f\big).$$

In the context of optimal transport, we shall be led to consider Hessian operators for functions f that are not of class C^2, and not even continuously differentiable. However, ∇f and $\nabla^2 f(x)$ will still be well-defined almost everywhere, and this will be sufficient to conduct the proofs. Here I should explain what it means for a function defined almost everywhere to be differentiable. Let ξ be a vector field defined on a domain of a neighborhood U of x; when y is close enough to x, there is a unique velocity $w \in T_xM$ such that $y = \gamma_1$, where γ is the constant-speed geodesic starting from x with initial velocity w; for simplicity I shall write $w = y - x$ (to be understood as $y = \exp_x w$). Then ξ is said to be covariantly differentiable at x in the direction v, if

$$\nabla_v \xi(x) := \lim_{y\to x;\ \frac{y-x}{|y-x|}\to\frac{v}{|v|}} |v|\,\left(\frac{\theta_{y\to x}\xi(y) - \xi(x)}{|y-x|}\right) \tag{14.5}$$

exists, where y varies on the domain of definition of ξ, and $\theta_{y \to x}$ is the parallel transport along the geodesic joining y to x. If ξ is defined everywhere in a neighborhood of x, then this is just the usual notion of covariant derivation. Formulas for (14.5) in coordinates are just the same as in the smooth case.

The following theorem is the main result of second differentiability for nonsmooth functions:

Theorem 14.1 (Second differentiability of semi-convex functions). *Let M be a smooth Riemannian manifold equipped with its volume measure, let U be an open subset of M, and let $\psi : U \to \mathbb{R}$ be locally semi-convex with a quadratic modulus of semi-convexity, in the sense of Definition 10.10. Then, for almost every $x \in U$, ψ is differentiable at x and there exists a symmetric operator $A : T_x M \to T_x M$, characterized by any one of the two equivalent properties*

(i) For any $v \in T_x M$, $\nabla_v (\nabla \psi)(x) = Av$;

(ii) $\psi(\exp_x v) = \psi(x) + \langle \nabla \psi(x), v \rangle + \dfrac{\langle A \cdot v, v \rangle}{2} + o(|v|^2)$ as $v \to 0$.

The operator A is denoted by $\nabla^2 \psi(x)$ and called the Hessian of ψ at x. When no confusion is possible, the quadratic form defined by A is also called the Hessian of ψ at x.

The trace of A is denoted by $\Delta \psi$ and called the Laplacian of ψ at x. The function $x \to \Delta \psi(x)$ coincides with the density of the absolutely continuous part of the distributional Laplacian of ψ.

Remark 14.2. The particular case when ψ is a convex function $\mathbb{R}^n \to \mathbb{R}$ is known as **Alexandrov's second differentiability theorem**. By extension, I shall use the terminology "Alexandrov's theorem" for the general statement where M is an arbitrary manifold. This theorem is more often stated in terms of Property (ii) than in terms of Property (i); but it is the latter that will be most useful for our purposes.

Remark 14.3. As the proof will show, Property (i) can be replaced by the more precise statement involving the subdifferential of ψ: If ξ is any vector field valued in $\nabla^- \psi$ (i.e. $\xi(y) \in \nabla^- \psi(y)$ for all y), then $\nabla_v \xi(x) = Av$.

Remark 14.4. For the main part of this course, we shall not need the full strength of Theorem 14.1, but just the particular case when ψ is continuously differentiable and $\nabla \psi$ is Lipschitz; then the proof becomes much simpler, and $\nabla \psi$ is almost everywhere differentiable in the usual sense. Still, on some occasions we shall need the full generality of Theorem 14.1.

Beginning of proof of Theorem 14.1. Since the notion of local semi-convexity with quadratic modulus is invariant by C^2 diffeomorphism, it is sufficient to prove it when $M = \mathbb{R}^n$. But a semi-convex function in $U \subset \mathbb{R}^n$ is just the sum of a quadratic form and a locally convex function (that is, a function which is convex in any convex subset of U). So it is actually sufficient to consider the special case when ψ is a convex function in a convex subset of $\mathbb{R}^n$. Then if $x \in U$ and B is a closed ball around x, included in U, let ψ_B be the restriction of ψ to B; since ψ is Lipschitz and convex, it can be extended into a Lipschitz convex function on the whole of $\mathbb{R}^n$ (take for instance the supremum of all supporting hyperplanes for ψ_B). In short, to prove Theorem 14.1 it is sufficient to treat the special case of a convex function $\psi : \mathbb{R}^n \to \mathbb{R}$. At this point the argument does not involve any more Riemannian geometry, but only convex analysis; so I shall postpone it to the Appendix (Theorem 14.23). □

The Jacobian determinant of the exponential map

Let M be a Riemannian manifold, and let ξ be a vector field on M (so for each x, $\xi(x)$ lies in T_xM). Recall the definition of the exponential map $T = \exp \xi$: Start from point x a geodesic curve with initial velocity $\xi(x) \in T_xM$, and follow it up to time 1 (it is not required that the geodesic be minimizing all along); the position at time 1 is denoted by $\exp_x \xi(x)$. As a trivial example, in Euclidean space, $\exp_x \xi(x) = x + \xi(x)$.

The computation of the Jacobian determinant of such a map is a classical exercise in Riemannian geometry, whose solution involves the Ricci curvature. One can take this computation as a theorem about the Ricci curvature (previously defined in terms of sectional or Riemann curvature), or as the mere definition of the Ricci curvature.

So let x be given, and let ξ be a vector field defined in a neighborhood of x, or almost everywhere in a neighborhood of x. Let $e_1, \ldots, e_n$ be an orthonormal basis of T_xM, and consider small variations of x in these directions $e_1, \ldots, e_n$, denoted abusively by $x + \delta e_1, \ldots, x + \delta e_n$. (Here $x + \delta e_j$ should be understood as, say, $\exp_x(\delta e_j)$; but it might also be any path $x(\delta)$ with $\dot{x}(0) = e_i$.) As $\delta \to 0$, the infinitesimal parallelepiped P_δ built on $(x + \delta e_1, \ldots, x + \delta e_n)$ has volume $\mathrm{vol}\,[P_\delta] \simeq \delta^n$. (It is easy to make sense of that by using local charts.) The quantity of interest is

$$\mathcal{J}(x) := \lim \frac{\mathrm{vol}\,[T(P_\delta)]}{\mathrm{vol}\,[P_\delta]}.$$

For that purpose, $T(P_\delta)$ can be approximated by the infinitesimal parallelogram built on $T(x + \delta e_1), \ldots, T(x + \delta e_n)$. Explicitly,

$$T(x + \delta e_i) = \exp_{x+\delta e_i}(\xi(x + \delta e_i)).$$

(If ξ is not defined at $x + \delta e_i$ it is always possible to make an infinitesimal perturbation and replace $x + \delta e_i$ by a point which is extremely close and at which ξ is well-defined. Let me skip this nonessential subtlety.)

Assume for a moment that we are in $\mathbb{R}^n$, so $T(x) = x + \xi(x)$. Then, by a classical result in real analysis, $\mathcal{J}(x) = |\det(\nabla T)| = |\det(I_n + \nabla \xi(x))|$. But in the genuinely Riemannian case, things are much more intricate (unless $\xi(x) = 0$) because the measurement of infinitesimal volumes changes as we move along the geodesic path $\gamma(t, x) = \exp_x(t\xi(x))$.

To appreciate this continuous change, let us parallel transport along the geodesic γ to define a new family $\mathbf{E}(t) = (e_1(t), \ldots, e_n(t))$ in $T_{\gamma(t)}M$. Since $(d/dt)\langle e_i(t), e_j(t)\rangle = \langle \dot{e}_i(t), e_j(t)\rangle + \langle e_i(t), \dot{e}_j(t)\rangle = 0$, the basis $\mathbf{E}(t)$ is an orthonormal basis of $T_{\gamma(t)}M$ for all t. (Here the dot symbol stands for the covariant derivation along γ.) Moreover, $e_1(t) = \dot{\gamma}(t, x)/|\dot{\gamma}(t, x)|$.

To express the Jacobian of the map $T = \exp \xi$, it will be convenient to consider the whole collection of maps $T_t = \exp(t\xi)$. For brevity, let us write

$$T_t(x + \delta \mathbf{E}) = \Big(T_t(x + \delta e_1), \ldots, T_t(x + \delta e_n)\Big);$$

then

$$T_t(x + \delta \mathbf{E}) \simeq T_t(x) + \delta\, \mathbf{J},$$

where

$$\mathbf{J} = (J_1, \ldots, J_n); \qquad J_i(t, x) := \left.\frac{d}{d\delta}\right|_{\delta=0} T_t(x + \delta e_i).$$

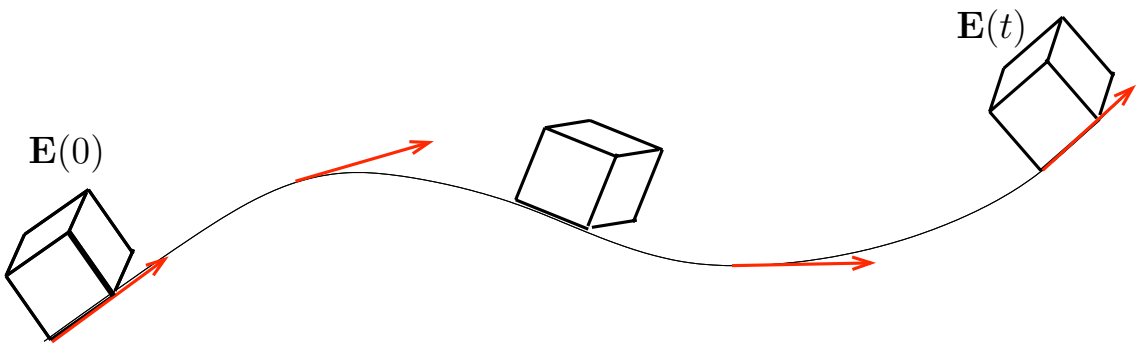


Fig. 14.2. The orthonormal basis **E**, here represented by a small cube, goes along the geodesic by parallel transport.

The vector fields J_i have been obtained by differentiating a family of geodesics depending on a parameter (here δ); such vector fields are called **Jacobi fields** and they satisfy a characteristic linear second-order equation known as the *Jacobi equation*. To write this equation, it will be convenient to express $J_1, \dots, J_n$ in terms of the basis $e_1, \dots, e_n$; so let $J_{ij} = \langle J_i, e_j \rangle$ stand for the j^{th} component of J_i in this basis. Then the matrix $J = (J_{ij})_{1 \le i,j \le n}$ satisfies the differential equation

$$\ddot{J}(t) + R(t) J(t) = 0, \tag{14.6}$$

where $R(t)$ is a matrix which depends on the Riemannian structure at $\gamma(t)$, and can be expressed in terms of the Riemann curvature tensor:

$$R_{ij}(t) = \Big\langle \mathrm{Riem}_{\gamma(t)}(\dot\gamma(t), e_i(t))\, \dot\gamma(t),\ e_j(t) \Big\rangle_{\gamma(t)}. \tag{14.7}$$

(All of these quantities depend implicitly on the starting point x.) The reader who prefers to stay away from the Riemann curvature tensor can take (14.6) as the equation defining the matrix R; the only things that one should know about this matrix are that (a) $R(t)$ is symmetric; (b) the first row of $R(t)$ vanishes (which is the same, modulo identification, as $R(t)\dot\gamma(t) = 0$); (c) tr $R(t) = \mathrm{Ric}_{\gamma_t}(\dot\gamma_t, \dot\gamma_t)$ (which one can also adopt as a definition of the Ricci tensor); (d) $R(t)$ is invariant under the transform $t \to 1 - t$, $E(t) \to -E(1-t)$, $\gamma_t \to \gamma_{1-t}$.

Equation (14.6) is of second order in time, and so it should come with initial conditions for both $J(0)$ and $\dot J(0)$. On one hand, since $T_0(y) = y$, one has

$$J_i(0) = \frac{d}{d\delta}\bigg|_{\delta=0} (x + \delta e_i) = e_i,$$

so $J(0)$ is just the identity matrix. On the other hand,

$$\begin{aligned} \dot J_i(0) = \frac{D}{Dt}\bigg|_{t=0} \frac{d}{d\delta}\bigg|_{\delta=0} T_t(x + \delta e_i) = \frac{D}{D\delta}\bigg|_{\delta=0} \frac{d}{dt}\bigg|_{t=0} T_t(x + \delta e_i) \\ = \frac{D}{D\delta}\bigg|_{\delta=0} \xi(x + \delta e_i) = (\nabla \xi) e_i, \end{aligned}$$

where $\nabla\xi$ is the covariant gradient of ξ. (It is easy to justify the exchange of derivatives by using the differentiability of ξ at x and the C^∞ regularity of $(t, y, \xi) \to \exp_y(t\xi)$.) So

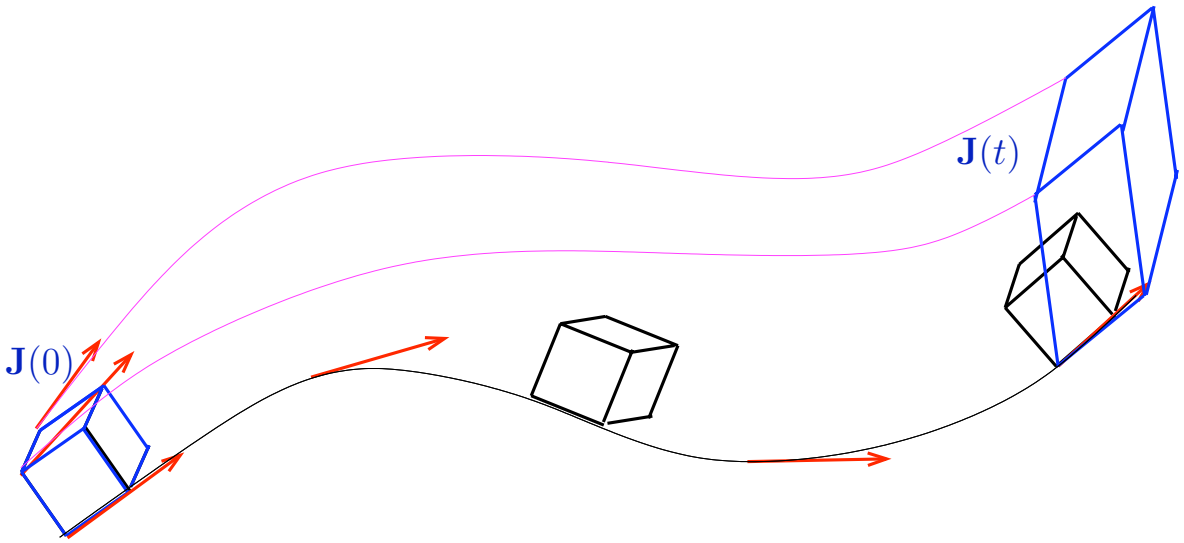


Fig. 14.3. At time $t = 0$, the matrices $\mathbf{J}(t)$ and $\mathbf{E}(t)$ coincide, but at later times they (may) differ, due to geodesic distortion.

$$\frac{d}{dt} J_{ij} = \frac{d}{dt} \langle J_i, e_j \rangle = \Big\langle \frac{DJ_i}{Dt}, e_j \Big\rangle = \langle (\nabla \xi) e_i, e_j \rangle.$$

We conclude that the initial conditions are

$$J(0) = I_n, \qquad \dot{J}(0) = \nabla \xi(x), \tag{14.8}$$

where in the second expression the linear operator $\nabla \xi(x)$ is identified with its matrix in the basis $\mathbf{E}$: $(\nabla \xi)_{ij} = \langle (\nabla \xi) e_i, e_j \rangle = \langle e_i \cdot \nabla \xi, e_j \rangle$. (Be careful, this is the contrary of the usual convention $A_{ij} = \langle A e_j, e_i \rangle$; anyway, later we shall work with symmetric operators, so it will not matter.)

From this point on, the problem is about a path $J(t)$ valued in the space $M_n(\mathbb{R})$ of real $n \times n$ matrices, and we can forget about the geometry: Parallel transport has provided a consistent identification of all the tangent spaces $T_{\gamma(t)} M$ with $\mathbb{R}^n$. This path depends on x via the initial conditions (14.8), so in the sequel we shall put that dependence explicitly. It might be very rough as a function of x, but it is very smooth as a function of t. The Jacobian of the map T_t is defined by

$$\mathcal{J}(t, x) = \det J(t, x),$$

and the formula for the differential of the determinant yields

$$\dot{\mathcal{J}}(t, x) = \mathcal{J}(t, x) \ \mathrm{tr} \left(\dot{J}(t, x) \, J(t, x)^{-1} \right), \tag{14.9}$$

at least as long as $J(t, x)$ is invertible (let's forget about that problem for the moment).

So it is natural to set

$$U := \dot{J} \, J^{-1}, \tag{14.10}$$

and to look for an equation on U. By differentiating (14.10) and using (14.6), we discover that

$$\dot{U} = \ddot{J} J^{-1} - \dot{J} J^{-1} \dot{J} J^{-1} = -R - U^2$$

(note that J and $\dot{J}$ do not necessarily commute). So the change of variables (14.10) has turned the second-order equation (14.6) into the *first-order* equation

$$\dot{U} + U^2 + R = 0, \tag{14.11}$$

which is of Ricatti type, that is, with a quadratic nonlinearity.

By taking the trace of (14.11), we arrive at

$$\frac{d}{dt}(\mathrm{tr}\ U) + \mathrm{tr}\,(U^2) + \mathrm{tr}\ R = 0.$$

Now the trace of $R(t,x)$ only depends on γ_t and $\dot{\gamma}_t$; in fact, as noticed before, it is precisely the value of the Ricci curvature at $\gamma(t)$, evaluated in the direction $\dot{\gamma}(t)$. So we have arrived at our first important equation involving Ricci curvature:

$$\frac{d}{dt}(\mathrm{tr}\ U) + \mathrm{tr}\,(U^2) + \mathrm{Ric}(\dot{\gamma}) = 0, \tag{14.12}$$

where of course $\mathrm{Ric}(\dot{\gamma})$ is an abbreviation for $\mathrm{Ric}_{\gamma(t)}(\dot{\gamma}(t), \dot{\gamma}(t))$.

Equation (14.12) holds true for any vector field ξ, as long as ξ is covariantly differentiable at x. But in the sequel, I shall only apply it in the particular case when ξ derives from a function: $\xi = \nabla\psi$; and ψ is locally semi-convex with a quadratic modulus of semi-convexity. There are three reasons for this restriction:

(a) Only such maps arise in optimal transport;

(b) Semi-convexity of ψ guarantees the almost everywhere differentiability of $\nabla\psi$, by Theorem 14.1;

(c) If $\xi = \nabla\psi$, then $\nabla\xi(x) = \nabla^2\psi(x)$ is *symmetric* and this will imply the symmetry of $U(t,x)$ at all times; this symmetry will allow to derive from (14.12) a closed inequality on $\mathrm{tr}\ U(t,x) = \mathcal{J}(t,x)$.

So from now on, $\xi = \nabla\psi$, where ψ is semi-convex. To prove the symmetry of $U(t,x)$, note that $U(0,x) = I_n$ and $\dot{U}(0,x) = \nabla^2\psi(x)$ (modulo identification) are symmetric, and the time-dependent matrix $R(t,x)$ is also symmetric, so $U(t,x)$ and its transpose $U(t,x)^*$ solve the same differential equation, with the same initial conditions. Then, by the uniqueness statement in the Cauchy–Lipschitz theorem, they have to coincide at all times where they are defined.

Inequality (14.12) cannot be recast as a differential equation involving only the Jacobian determinant (or equivalently $\mathrm{tr}\ U(t,x)$), since the quantity $\mathrm{tr}\,(U^2)$ in (14.12) cannot be expressed in terms of $\mathrm{tr}\ U$. However, the symmetry of U allows to use the Cauchy–Schwarz inequality, in the form

$$\mathrm{tr}\,(U^2) \geq \frac{(\mathrm{tr}\ U)^2}{n};$$

then, by plugging this inequality into (14.12), we obtain an important differential inequality involving Ricci curvature:

$$\frac{d}{dt}(\mathrm{tr}\ U) + \frac{(\mathrm{tr}\ U)^2}{n} + \mathrm{Ric}(\dot{\gamma}) \leq 0. \tag{14.13}$$

There are several ways to rewrite this result in terms of the Jacobian $\mathcal{J}(t)$. For instance, by differentiating the formula

$$\mathrm{tr}\ U = \frac{\dot{\mathcal{J}}}{\mathcal{J}},$$

one obtains easily

$$\frac{d}{dt}(\mathrm{tr}\ U) + \frac{(\mathrm{tr}\ U)^2}{n} = \frac{\ddot{\mathcal{J}}}{\mathcal{J}} - \left(1 - \frac{1}{n}\right)\left(\frac{\dot{\mathcal{J}}}{\mathcal{J}}\right)^2.$$

So (14.13) becomes

$$\frac{\ddot{\mathcal{J}}}{\mathcal{J}} - \left(1 - \frac{1}{n}\right)\left(\frac{\dot{\mathcal{J}}}{\mathcal{J}}\right)^2 \leq -\operatorname{Ric}(\dot\gamma). \tag{14.14}$$

For later purposes, a convenient formulation consists in defining $\mathcal{D}(t) := \mathcal{J}(t)^{1/n}$ (which one can think of as a coefficient of mean distortion), and then the left-hand side of (14.14) is exactly $n\ddot{\mathcal{D}}/\mathcal{D}$. So

$$\frac{\ddot{\mathcal{D}}}{\mathcal{D}} \leq -\frac{\operatorname{Ric}(\dot\gamma)}{n}. \tag{14.15}$$

Another useful formula is obtained by considering $\ell(t) := -\log \mathcal{J}(t)$, and then (14.13) writes

$$\ddot\ell(t) \geq \frac{\dot\ell(t)^2}{n} + \operatorname{Ric}(\dot\gamma). \tag{14.16}$$

In all these formulas, we have always taken the point $t = 0$ as the starting time, but it is clear that we could do just the same with any starting time $t_0 \in [0,1]$, that is, consider, instead of $T_t(x) = \exp(t\nabla\psi(x))$, the map $T_{t_0\to t}(x) = \exp((t-t_0)\nabla\psi(x))$. Then all the differential inequalities are unchanged; the only difference is that the Jacobian determinant at time $t = 0$ is not necessarily 1.

Taking out the direction of motion

The previous formulas are quite sufficient to derive many useful geometric consequences. However, one can refine them by taking advantage of the fact that *curvature is not felt in the direction of motion*. In other words, if one is travelling along some geodesic γ, one will never be able to detect some curvature by considering variations (in the initial position, or initial velocity) in the direction of γ itself: the path will always be the same, up to reparametrization. This corresponds to the property $R(t)\dot\gamma(t) = 0$, where $R(t)$ is the matrix appearing in (14.6). In short, curvature is felt only in $n-1$ directions out of n. This loose principle often leads to a refinement of estimates by a factor $(n-1)/n$.

Here is a recipe to "separate out" the direction of motion from the rest. As before, assume that the first vector of the orthonormal basis $\mathbf{J}(0)$ is $e_1(0) = \dot\gamma(0)/|\dot\gamma(0)|$ (the case when $\dot\gamma(0) = 0$ can be treated separately). Set $u_{/\!/} = u_{11}$ (this is the coefficient in U which corresponds to the direction of motion only), and define $U_\perp$ as the $(n-1)\times(n-1)$ matrix obtained by removing the first line and first column in U. Of course, $\operatorname{tr}(U) = u_{/\!/} + \operatorname{tr}(U_\perp)$. Next decompose the Jacobian determinant $\mathcal{J}$ into a parallel and an orthogonal contributions:

$$\mathcal{J} = \mathcal{J}_{/\!/}\mathcal{J}_\perp, \qquad \mathcal{J}_{/\!/}(t) = \exp\left(\int_0^t u_{/\!/}(s)\,ds\right).$$

Further define parallel and orthogonal distortions by

$$\mathcal{D}_{/\!/} = \mathcal{J}_{/\!/}, \qquad \mathcal{D}_\perp = \mathcal{J}_\perp^{\frac{1}{n-1}};$$

and, of course,

$$\ell_{/\!/} = -\log \mathcal{J}_{/\!/}, \qquad \ell_\perp = -\log \mathcal{J}_\perp. \tag{14.17}$$

Now, since the first row of $R(t)$ vanishes, the equation (14.11) implies

$$\dot u_{/\!/} = -\sum_j u_{1j}^2 \leq -u_{11}^2 = -u_{/\!/}^2.$$

It follows easily that

$$\ddot{\ell}_{/\!/} \geq \dot{\ell}_{/\!/}^2, \tag{14.18}$$

or equivalently

$$\ddot{\mathcal{J}}_{/\!/} \leq 0, \tag{14.19}$$

so $\mathcal{D}_{/\!/} = \mathcal{J}_{/\!/}$ is always a concave function of t, independently of the curvature of M, and the same holds true of course for $\mathcal{D}_{/\!/}$ which coincides with $\mathcal{J}_{/\!/}$.

Now let us take care of the orthogonal part: By putting together (14.9), (14.10), (14.11), (14.18), it is immediate that

$$\ddot{\ell}_\perp = -\frac{d}{dt}(\mathrm{tr}\ U) - \ddot{\ell}_{/\!/} = \mathrm{tr}\,(U^2) + \mathrm{Ric}(\dot\gamma) - \sum u_{1j}^2.$$

Since $\mathrm{tr}\ U^2 = \mathrm{tr}\,(U_\perp)^2 + 2\sum u_{1j}^2$, it follows that

$$\ddot{\ell}_\perp \geq \mathrm{tr}\,(U_\perp^2) + \mathrm{Ric}(\dot\gamma). \tag{14.20}$$

Then in the same manner as before, one can obtain

$$\ddot{\ell}_\perp \geq \frac{(\dot\ell_\perp)^2}{n-1} + \mathrm{Ric}(\dot\gamma), \tag{14.21}$$

$$\frac{\ddot{\mathcal{D}}_\perp}{\mathcal{D}_\perp} \leq -\frac{\mathrm{Ric}(\dot\gamma)}{n-1}. \tag{14.22}$$

To summarize: The basic inequalities for $\ell_\perp$ and $\ell_{/\!/}$ are the same as for ℓ, but with the exponent n replaced by $n-1$ in the case of $\ell_\perp$, and 1 in the case of $\ell_{/\!/}$; and the number $\mathrm{Ric}(\dot\gamma)$ replaced by 0 in the case of $\ell_{/\!/}$.

Positivity of the Jacobian

Unlike the distance function, the exponential map is always smooth. But this does not prevent the Jacobian determinant $\mathcal{J}(t)$ to vanish, i.e. the matrix $J(t)$ to become singular (not invertible). Then computations such as (14.9) break down. So all the computations performed before are only valid if $\mathcal{J}(t)$ is positive for all $t \in (0,1)$.

In terms of the quantity $\ell(t) = -\log \mathcal{J}(t)$, the vanishing of the Jacobian determinant corresponds to a divergence $\ell(t) \to \infty$. Readers familiar with ordinary differential equations will have no trouble believing that these events are not rare: Indeed, ℓ solves a Ricatti-type equation such as (14.16), and such equations often lead to blow-up in finite time. For instance, consider a function $\ell(t)$ that solves

$$\ddot\ell \geq \frac{(\dot\ell)^2}{n-1} + K,$$

where $K > 0$. Consider a time t_0 where ℓ is minimum, so $\dot\ell(t_0) = 0$. Then, ℓ cannot be defined on a time-interval larger than $[t_0 - T, t_0 + T]$, where $T := \pi\sqrt{(n-1)/K}$. So the Jacobian has to vanish at some time, and we even have a bound on this time. (With a bit more work, this estimate implies the *Bonnet–Myers theorem*, which asserts that the diameter of M cannot be larger than $\pi\sqrt{(n-1)/K}$ if $\mathrm{Ric} \geq K\,g$.)

The vanishing of the Jacobian may occur even along geodesics that are minimizing for all times: Consider for instance $\xi(x) = -2x$ in $\mathbb{R}^n$; then the image of $\exp(t\xi)$ is reduced to a single point when $t = 1/2$. However, in the case of optimal transport, the Jacobian cannot vanish at intermediate times, at least for almost all initial points: Recall indeed the last part of Theorem 11.3. This property can be seen as a result of the very special choice of the velocity field ξ, which is the gradient *of a $d^2/2$-convex function*; or as a consequence of the "no-crossing" property explored in Chapter 8.

Bochner's formula

So far, we have discussed curvature from a *Lagrangian* point of view, that is, by going along a geodesic path $\gamma(t)$, keeping the memory of the initial position. It is useful to be also familiar with the **Eulerian** point of view, in which the focus is not on the trajectory, but on the velocity field $\xi = \xi(t,x)$. To switch from Lagrangian to Eulerian description, just write

$$\dot\gamma(t) = \xi\big(t, \gamma(t)\big). \tag{14.23}$$

In general, this can be a subtle issue because two trajectories might cross, and then there would be no way to define a meaningful velocity field $\xi(t,\cdot)$ at the crossing point. However, if a smooth vector field $\xi = \xi(0,\cdot)$ is given, then around a given point x_0 the trajectories $\gamma(t,x) = \exp(t\xi(x))$ do not cross in short time, and then one can define $\xi(t,x)$ without ambiguity. The covariant differentiation of (14.23) along ξ itself, and the geodesic equation $\ddot\gamma = 0$, yield

$$\frac{\partial \xi}{\partial t} + \nabla_\xi \xi = 0, \tag{14.24}$$

which is the **pressureless Euler equation**. From a physical point of view, this equation describes the velocity field of a bunch of particles which travel along geodesic curves without interacting. The derivation of (14.24) will fail when the geodesic paths start to cross, at which point the solution to (14.24) would typically lose smoothness and need reinterpretation. But for the sequel, we only need (14.24) to be satisfied in short time, and locally around x.

Now, all the discussion about Ricci curvature can be recast in Eulerian terms. Let $\gamma(t,x) = \exp_x(t\xi(x))$; by the definition of the covariant gradient, we have

$$\dot{\mathbf{J}}(t,x) = \nabla\xi(t,\gamma(t,x))\,\mathbf{J}(t,x)$$

(the same formula that we had before at time $t = 0$). Under the identification of $\mathbb{R}^n$ with $T_{\gamma(t)}M$ provided by the basis $\mathbf{E}(t)$, we can identify $\mathbf{J}$ with the matrix J, and then

$$U(t,x) = \dot J(t,x)\, J(t,x)^{-1} = \nabla\xi\big(t,\gamma(t,x)\big), \tag{14.25}$$

where again the linear operator $\nabla\xi$ is identified with its matrix in the basis provided by $\mathbf{E}$.

Then $\mathrm{tr}\, U(t,x) = \mathrm{tr}\, \nabla\xi(t,x)$ coincides with the divergence of $\xi(t,\cdot)$, evaluated at x. By the chain rule and (14.24),

$$\begin{aligned}
\frac{d}{dt}(\mathrm{tr}\, U)(t,x) &= \frac{d}{dt}(\nabla\cdot\xi)(t,\gamma(t,x)) \\
&= \nabla\cdot\left(\frac{\partial\xi}{\partial t}(t,\gamma(t,x)\right) + \dot\gamma(t,x)\cdot\nabla(\nabla\cdot\xi)(t,\gamma(t,x)) \\
&= \Big(-\nabla\cdot(\nabla_\xi\xi) + \xi\cdot\nabla(\nabla\cdot\xi)\Big)(t,\gamma(t,x)).
\end{aligned}$$

So the Lagrangian formula (14.12) can be translated into the Eulerian formula

$$-\nabla\cdot(\nabla_\xi\xi) + \xi\cdot\nabla(\nabla\cdot\xi) + \mathrm{tr}\,(\nabla\xi)^2 + \mathrm{Ric}(\xi) = 0. \tag{14.26}$$

All functions here are evaluated at $(t,\gamma(t,x))$, and of course we can choose $t = 0$, and x arbitrary. So (14.26) is an identity that holds true for any smooth (say C^2) vector

field ξ on our manifold M. Of course it can also be established directly by a coordinate computation.[1]

While formula (14.26) holds true for all vector field ξ, if $\nabla\xi$ is symmetric then two simplifications arise:

(a) $\nabla_\xi \xi = \nabla\xi \cdot \xi = \nabla \cdot \dfrac{|\xi|^2}{2}$;

(b) $\mathrm{tr}\,(\nabla\xi)^2 = \|\nabla\xi\|^2_{\mathrm{HS}}$, where HS stands for the Hilbert-Schmidt norm.

Then (14.26) becomes

$$-\Delta\frac{|\xi|^2}{2} + \xi\cdot\nabla(\nabla\cdot\xi) + \|\nabla\xi\|^2_{\mathrm{HS}} + \mathrm{Ric}(\xi) = 0. \tag{14.27}$$

We shall apply it only in the case when ξ is a gradient: $\xi = \nabla\psi$; then $\nabla\xi = \nabla^2\psi$ is indeed symmetric, and the resulting formula is

$$-\Delta\frac{|\nabla\psi|^2}{2} + \nabla\psi\cdot\nabla(\Delta\psi) + \|\nabla^2\psi\|^2_{\mathrm{HS}} + \mathrm{Ric}(\nabla\psi) = 0. \tag{14.28}$$

The identity (14.26), or its particular case (14.28), is called the **Bochner–Weitzenböck–Lichnérowicz formula**, or just Bochner's formula.[2]

Remark 14.5. With the ansatz $\xi = \nabla\psi$, the pressureless Euler equation (14.24) reduces to the **Hamilton–Jacobi equation**

$$\frac{\partial\psi}{\partial t} + \frac{|\nabla\psi|^2}{2} = 0. \tag{14.29}$$

One can use this equation to obtain (14.28) directly, instead of first deriving (14.26). Here equation (14.29) is to be understood in viscosity sense (otherwise there are many spurious solutions); in fact the reader might just as well take the identity

$$\psi(t,x) = \inf_{y\in M}\left[\psi(y) + \frac{d(x,y)^2}{2t}\right]$$

as the definition of the solution of (14.29). Then the geodesic curves γ starting with $\gamma(0) = x$, $\dot\gamma(0) = \nabla\psi(x)$ are called *characteristic curves* of the equation (14.29).

Remark 14.6. Here I have not tried to derive Bochner's formula for nonsmooth functions. This could be done for semi-convex ψ, with an appropriate "compensated" definition for $-\Delta\frac{|\nabla\psi|^2}{2} + \nabla\psi\cdot\nabla(\Delta\psi)$. In fact, the semi-convexity of $\nabla\psi$ prevents the formation of instantaneous shocks, and will allow the Lagrangian/Eulerian duality for a short time.

Remark 14.7. The operator $U(t,x)$ coincides with $\nabla^2\psi(t,\gamma(t,x))$, which is another way to see that it is symmetric for $t > 0$.

[1] With the notation $\nabla_\xi = \xi\cdot\nabla$ (which is classical in fluid mechanics), and $\mathrm{tr}\,(\nabla\xi)^2 = \nabla\xi\cdot\cdot\nabla\xi$, formula (14.26) takes the amusing form $-\nabla\cdot\xi\cdot\nabla\xi + \xi\cdot\nabla\nabla\cdot\xi + \nabla\xi\cdot\cdot\nabla\xi + \mathrm{Ric}(\xi) = 0$.

[2] In (14.26) or (14.28) I have written Bochner's formula in purely "metric" terms, which will probably look quite ugly to many geometer readers. An equivalent but more "topological" way to write Bochner's formula is

$$\Delta = -\nabla\nabla^* + \mathrm{Ric},$$

where $\Delta = -(dd^* + d^*d)$ is the Laplace operator on 1-forms, ∇ is the covariant differentiation (under the identification of a 1-form with a vector field) and the adjoints are in $L^2(\mathrm{vol}\,)$. Also I should note that the name "Bochner formula" is attributed to a number of related identities.

From that point on, we shall only work with (14.28). Of course, by using the Cauchy–Schwarz identity as before, we can bound below $\|\nabla^2\psi\|^2_{\mathrm{HS}}$ by $(\Delta\psi)^2/n$; therefore (14.25) implies

$$\Delta\frac{|\nabla\psi|^2}{2} - \nabla\psi\cdot\nabla(\Delta\psi) \geq \frac{(\Delta\psi)^2}{n} + \mathrm{Ric}(\nabla\psi). \tag{14.30}$$

Apart from regularity issues, this inequality is strictly equivalent to (14.13), and therefore to (14.14) or (14.15).

Not so much has been lost when going from (14.28) to (14.30): there is still equality in (14.30) at all points x where $\nabla^2\psi(x)$ is a multiple of the identity.

It is also possible to take out the direction of motion, $\widehat{\nabla\psi} := (\nabla\psi)/|\nabla\psi|$, from the Bochner identity. The Hamilton–Jacobi equation implies $\partial_t\widehat{\nabla\psi} + \nabla^2\psi\cdot\widehat{\nabla\psi} = 0$, so

$$\partial_t\Big\langle \nabla^2\psi\cdot\widehat{\nabla\psi}, \widehat{\nabla\psi}\Big\rangle = -\big\langle\nabla^2(|\nabla\psi|^2/2)\cdot\widehat{\nabla\psi},\widehat{\nabla\psi}\big\rangle - 2\,\big\langle\nabla^2\psi\cdot(\nabla^2\psi\cdot\widehat{\nabla\psi}),\widehat{\nabla\psi}\big\rangle,$$

and by symmetry the latter term can be rewritten $-2\,|(\nabla^2\psi)\cdot\widehat{\nabla\psi}|^2$. From this one easily obtains the following refinement of Bochner's formula: Define

$$\Delta_{/\!/} f = \big\langle\nabla^2 f\cdot\widehat{\nabla\psi},\,\widehat{\nabla\psi}\big\rangle, \qquad \Delta_\perp = \Delta - \Delta_{/\!/},$$

then

$$\begin{cases} \Delta_{/\!/}\frac{|\nabla\psi|^2}{2} - \nabla\psi\cdot\nabla\Delta_{/\!/}\psi + 2\big|(\nabla^2\psi)\cdot\widehat{\nabla\psi}\big|^2 \geq (\Delta_{/\!/}\psi)^2 \\ \\ \Delta_\perp\frac{|\nabla\psi|^2}{2} - \nabla\psi\cdot\nabla\Delta_\perp\psi - 2\big|(\nabla^2\psi)\cdot\widehat{\nabla\psi}\big|^2 \geq \|\nabla^2_\perp\psi\|^2_{\mathrm{HS}} + \mathrm{Ric}(\nabla\psi). \end{cases} \tag{14.31}$$

This is the "Bochner formula with the direction of motion taken out". I have to confess that I never saw these frightening formulas anywhere, and don't know whether they have any use. But of course, they are equivalent to their Lagrangian counterpart, that will play a crucial role in the sequel.

Analytic and geometric consequences of Ricci curvature bounds

Inequalities (14.13), (14.14), (14.15) and (14.30) are the "working heart" of Ricci curvature analysis. Many gometric and analytic consequences follow from these estimates.

Here is a first example coming from analysis and partial differential equations theory: If the Ricci curvature of M is globally bounded below ($\inf_x \mathrm{Ric}_x > -\infty$), then there exists a unique *heat kernel*, that is a measurable function $p_t(x,y)$ ($t>0$, $x\in M$, $y\in M$), integrable in y, smooth outside of the diagonal $x=y$, such that $f(t,x) := \int p_t(x,y)\,f_0(y)\,d\mathrm{vol}\,(y)$ solves the heat equation $\partial_t f = \Delta f$ with initial datum f_0.

Here is another example in which some topological information can be recovered from Ricci bounds: If M is a manifold with nonnegative Ricci curvature (for each x, $\mathrm{Ric}_x \geq 0$), and there exists a line in M, that is, a geodesic γ which is minimizing for *all* values of time $t\in\mathbb{R}$, then M is isometric to $\mathbb{R}\times M'$, for some Riemannian manifold M'. This is the **splitting theorem**, in a form proven by Cheeger and Gromoll.

Many *quantitative* statements can be obtained from (i) a lower bound on the Ricci curvature and (ii) an upper bound on the dimension of the manifold. Here below is a (grossly nonexhaustive) list of some famous such results. In the statements to come, M is always assumed to be a smooth complete Riemannian manifold, vol stands for the Riemannian volume on M, Δ for the Laplace operator and d for the Riemannian distance;

K is the lower bound on the Ricci curvature, and n is the dimension of M. Also, if A is a measurable set, then A^r will denote its r-neighborhood, that is the set of points that lie at distance at most r from A. Finally, the "model space" is the connected Riemannian manifold with constant sectional curvature which has the same dimension as M, and a Ricci curvature equal to K (more rigorously, Kg, where g is the metric tensor on the model space).

1. Volume growth estimates: The **Bishop–Gromov** inequality (also called Riemannian volume comparison theorem) states that the volume of balls does not increase faster than the volume of balls in the model space. In formulas: for all $x \in M$,

$$\frac{\mathrm{vol}\,[B_r(x)]}{V(r)} \quad \text{is a nonincreasing function of } r,$$

where

$$V(r) = \int_0^r S(r')\,dr', \qquad S(r) = c_{n,K} \begin{cases} \sin^{n-1}\left(\sqrt{\dfrac{K}{n-1}}\,s\right) & \text{if } K > 0 \\ s^{n-1} & \text{if } K = 0 \\ \sinh^{n-1}\left(\sqrt{\dfrac{|K|}{n-1}}\,s\right) & \text{if } K < 0. \end{cases}$$

Here of course $S(r)$ denotes the surface of $B_r(0)$ in the model space, that is the $(n-1)$-dimensional volume of $\partial B_r(0)$, and $c_{n,K}$ is a nonessential normalizing constant.

2. Diameter estimates: The **Bonnet–Myers theorem** states that, if $K > 0$, then M is compact and more precisely

$$\mathrm{diam}\,(M) \le \pi\sqrt{\frac{n-1}{K}},$$

with equality for the model sphere.

3. Spectral gap inequalities: If $K > 0$, then the spectral gap λ_1 of the nonnegative operator $-\Delta$ is bounded below:

$$\lambda_1 \ge \frac{nK}{n-1},$$

with equality again for the model sphere.

4. (Sharp) Sobolev inequalities: If $K > 0$ and $n \ge 2$, let $\mu = \mathrm{vol}\,/\mathrm{vol}\,[M]$ be the normalized volume measure on M; then for any smooth function on M,

$$\|f\|^2_{L^{2^\star}(\mu)} \le \|f\|^2_{L^2(\mu)} + \frac{4}{Kn(n-2)}\,\|\nabla f\|^2_{L^2(\mu)}, \qquad 2^\star = \frac{2n}{n-2},$$

and those constants are sharp for the model sphere.

5. Heat kernel bounds: There are many of them, in particular the celebrated **Li–Yau estimates**: If $K \ge 0$, then the heat kernel $p_t(x,y)$ satisfies

$$p_t(x,y) \le \frac{C}{\mathrm{vol}\,[B_{\sqrt{t}}(x)]}\exp\left(-\frac{d(x,y)^2}{2\,Ct}\right),$$

for some constant C which only depends on n. For $K < 0$, a similar bound holds true, only now C depends on K and there is an additional factor e^{Ct}. There are also pointwise estimates on the derivatives of $\log p_t$, in relation with **Harnack inequalities**.

The list could go on. More recently, Ricci curvature has been at the heart of Perelman's solution (so it seems) of the celebrated Poincaré conjecture, and more generally the topological classification of three-dimensional manifolds. Indeed, Perelman's argument is based on Hamilton's idea to use Ricci curvature in order to define a "heat flow" in the space of metrics, via the partial differential equation

$$\frac{\partial g}{\partial t} = -2\,\mathrm{Ric}(g), \tag{14.32}$$

where $\mathrm{Ric}(g)$ is the Ricci tensor associated with the metric g — which can be thought of as something like $-\Delta g$. The flow defined by (14.32) is called the Ricci flow. Some time ago, Hamilton had already used its properties to show that a compact simply connected three-dimensional Riemannian manifold with positive Ricci curvature is automatically diffeomorphic to the sphere S^3.

Change of reference measure and effective dimension

For various reasons, one is often led to consider a reference measure ν that is not the volume measure vol, but, say, $\nu(dx) = e^{-V(x)}\,\mathrm{vol}\,(dx)$, for some function $V : M \to \mathbb{R}$, which in this chapter will always be assumed to be of class C^2. The metric–measure space (M, d, ν), where d stands for the geodesic distance, may be of interest in its own right, or may appear as a limit of Riemannian manifolds, in a sense that will be studied in Part III of these notes.

Of course, such a change of reference measure affects Jacobian determinants; so Ricci curvature estimates will lose their geometric meaning unless one changes the definition of Ricci tensor to take the new reference measure into account. This might perturb the dependence of all the estimates on the dimension, so it might also be a good idea to introduce an "effective dimension" N, which will always be larger than the true dimension n.

The most well-known example may be the **Gaussian measure** in $\mathbb{R}^n$, which I shall denote by $\gamma^{(n)}$ (do not confuse it with a geodesic!). It is a matter of experience that most theorems which we encounter about the Gaussian measure can be written just the same in dimension 1 or in dimension n, or even in infinite dimension, when properly interpreted. In fact, the effective dimension of $(\mathbb{R}^n, \gamma^{(n)})$ is infinite, in a certain sense, whatever n. I admit that this perspective might look strange, and might be the result of lack of imagination; but in any case, it will fit very well into the picture (in terms of sharp constants for geometric inequalities, etc.)

So let again

$$T_t(x) = \gamma(t, x) = \exp_x\big(t\nabla\psi(x)\big);$$

now the Jacobian determinant is

$$\mathcal{J}(t,x) = \lim_{r\downarrow 0} \frac{\nu\big[T_t(B_r(x))\big]}{\nu[B_r(x)]} = \frac{e^{-V(T_t(x))}}{e^{-V(x)}}\ \mathcal{J}_0(t,x).$$

where $\mathcal{J}_0$ is the Jacobian corresponding to $V \equiv 0$ (that is, to $\nu = \mathrm{vol}$).

Then (with dots still standing again for derivation with respect to t),

$$(\log \mathcal{J})^{\cdot}(t,x) = (\log \mathcal{J}_0)^{\cdot}(t,x) - \dot\gamma(t,x)\cdot \nabla V(\gamma(t,x)),$$

$$(\log \mathcal{J})^{\cdot\cdot}(t,x) = (\log \mathcal{J}_0)^{\cdot\cdot}(t,x) - \Big\langle \nabla^2 V(\gamma(t,x))\cdot\dot\gamma(t,x),\, \dot\gamma(t,x)\Big\rangle.$$

For later purpose it will be useful to keep track of all error terms in the inequalities. So rewrite (14.12) as

$$(\mathrm{tr}\ U)^{\cdot} + \frac{(\mathrm{tr}\ U)^2}{n} + \mathrm{Ric}(\dot\gamma) = -\left\| U - \left(\frac{\mathrm{tr}\ U}{n}\right) I_n \right\|_{\mathrm{HS}}^2. \tag{14.33}$$

Then the left-hand side in (14.33) becomes

$$\begin{aligned}(\log \mathcal{J}_0)^{\cdot\cdot} + \frac{[(\log \mathcal{J}_0)^{\cdot}]^2}{n} + \mathrm{Ric}(\dot\gamma)& \\ = (\log \mathcal{J})^{\cdot\cdot} + \langle \nabla^2 V(\gamma)\cdot\dot\gamma,\dot\gamma\rangle + \frac{[(\log \mathcal{J})^{\cdot} + \dot\gamma\cdot\nabla V(\gamma)]^2}{n} + \mathrm{Ric}(\dot\gamma).&\end{aligned}$$

By using the identity

$$\frac{a^2}{n} = \frac{(a+b)^2}{N} - \frac{b^2}{N-n} + \frac{n}{N(N-n)}\left(b - a\,\frac{N-n}{n}\right)^2, \tag{14.34}$$

we see that

$$\begin{aligned}&\frac{\Big[(\log \mathcal{J})^{\cdot} + \dot\gamma\cdot\nabla V(\gamma)\Big]^2}{n} \\ &\qquad = \frac{\big[(\log \mathcal{J})^{\cdot}\big]^2}{N} - \frac{(\dot\gamma\cdot\nabla V(\gamma))^2}{N-n} + \frac{n}{N(N-n)}\left[\left(\frac{N-n}{n}\right)(\log \mathcal{J})^{\cdot} + \frac{N}{n}\dot\gamma\cdot\nabla V(\gamma)\right]^2 \\ &\qquad = \frac{\big(\log \mathcal{J})^{\cdot}\big]^2}{N} - \frac{(\dot\gamma\cdot\nabla V(\gamma))^2}{N-n} + \frac{n}{N(N-n)}\left[\frac{N-n}{n}(\log \mathcal{J}_0)^{\cdot} + \dot\gamma\cdot\nabla V(\gamma)\right]^2 \\ &\qquad = \frac{\big[(\log \mathcal{J})^{\cdot}\big]^2}{N} - \frac{(\dot\gamma\cdot\nabla V(\gamma))^2}{N-n} + \frac{n}{N(N-n)}\left[\frac{N-n}{n}\,\mathrm{tr}\ U + \dot\gamma\cdot\nabla V(\gamma)\right]^2\end{aligned}$$

To summarize these computations it will be useful to introduce some more notation: first, as usual, the negative logarithm of the Jacobian determinant:

$$\ell(t,x) := -\log \mathcal{J}(t,x); \tag{14.35}$$

and then, the **modified Ricci tensor**:

$$\mathrm{Ric}_{N,\nu} := \mathrm{Ric} + \nabla^2 V - \frac{\nabla V\otimes\nabla V}{N-n}, \tag{14.36}$$

where the tensor product $\nabla V\otimes\nabla V$ is a quadratic form on TM, defined by its action on tangent vectors as

$$\big(\nabla V\otimes\nabla V\big)_x(v) = (\nabla V(x)\cdot v)^2;$$

so

$$\mathrm{Ric}_{N,\nu}(\dot\gamma) = (\mathrm{Ric} + \nabla^2 V)(\dot\gamma) - \frac{(\nabla V\cdot\dot\gamma)^2}{N-n}.$$

Note that $\mathrm{Ric}_{\infty,\nu} = \mathrm{Ric} + \nabla^2 V$, while $\mathrm{Ric}_{n,\mathrm{vol}} = \mathrm{Ric}$.

The conclusion of the preceding computations is that

$$\ddot{\ell} = \frac{\dot{\ell}^2}{N} + \mathrm{Ric}_{N,\nu}(\dot\gamma) + \left\| U - \left(\frac{\mathrm{tr}\ U}{n}\right) I_n \right\|_{\mathrm{HS}}^2 + \frac{n}{N(N-n)} \left[\left(\frac{N-n}{n}\right) \mathrm{tr}\ U + \dot\gamma \cdot \nabla V(\gamma) \right]^2 \tag{14.37}$$

When $N = \infty$ this takes a simpler form:

$$\ddot{\ell} = \mathrm{Ric}_{\infty,\nu}(\dot\gamma) + \left\| U - \left(\frac{\mathrm{tr}\ U}{n}\right) I_n \right\|_{\mathrm{HS}}^2 \tag{14.38}$$

When $N < \infty$ one can introduce

$$\mathcal{D}(t) := \mathcal{J}(t)^{\frac{1}{N}},$$

and then formula (14.37) becomes

$$-N\frac{\ddot{\mathcal{D}}}{\mathcal{D}} = \mathrm{Ric}_{N,\nu}(\dot\gamma) + \left\| U - \left(\frac{\mathrm{tr}\ U}{n}\right) I_n \right\|_{\mathrm{HS}}^2 + \frac{n}{N(N-n)} \left[\left(\frac{N-n}{n}\right) \mathrm{tr}\ U + \dot\gamma \cdot \nabla V(\gamma) \right]^2 \tag{14.39}$$

Of course, it is a trivial corollary of (14.37) and (14.39) that

$$\begin{cases} \ddot{\ell} \geq \dfrac{\dot{\ell}^2}{N} + \mathrm{Ric}_{N,\nu}(\dot\gamma) \\ \\ -N\dfrac{\ddot{\mathcal{D}}}{\mathcal{D}} \geq \mathrm{Ric}_{N,\nu}(\dot\gamma). \end{cases} \tag{14.40}$$

Finally, if one wishes, one can also take out the direction of motion (skip at first reading and go directly to the next section). Define, with self-explicit notation,

$$\mathcal{J}_\perp(t,x) = \mathcal{J}_{0,\perp}(t,x)\, \frac{e^{-V(T_t(x))}}{e^{-V(x)}},$$

and $\ell_\perp = -\log \mathcal{J}_\perp$, $\mathcal{D}_\perp = \mathcal{J}_\perp^{\frac{1}{N}}$. Now, in place of (14.33), use

$$(\mathrm{tr}\ U_\perp)^{\cdot} + \frac{(\mathrm{tr}\ U_\perp)^2}{n-1} + \mathrm{Ric}(\dot\gamma) = -\left\| U_\perp - \left(\frac{\mathrm{tr}\ U_\perp}{n-1}\right) I_{n-1} \right\|_{\mathrm{HS}}^2 - \sum_{j=2}^n u_{1j}^2 \tag{14.41}$$

as a starting point. Then computations quite similar to the ones above lead to

$$\ddot{\ell}_\perp = \frac{(\dot{\ell}_\perp)^2}{N-1} + \mathrm{Ric}_{N,\nu}(\dot\gamma) + \left\| U_\perp - \left(\frac{\mathrm{tr}\ U_\perp}{n-1}\right) I_{n-1} \right\|_{\mathrm{HS}}^2 + \frac{n-1}{(N-1)(N-n)} \left[\left(\frac{N-n}{n-1}\right) \mathrm{tr}\ U + \dot\gamma \cdot \nabla V(\gamma) \right]^2 + \sum_{j=2}^n u_{1j}^2. \tag{14.42}$$

In the case $N = \infty$, this reduces to

$$\ddot{\ell}_\perp = \mathrm{Ric}_{\infty,\nu}(\dot\gamma) + \left\| U_\perp - \left(\frac{\mathrm{tr}\ U_\perp}{n-1}\right) I_{n-1} \right\|_{\mathrm{HS}}^2 + \sum_{j=2}^n u_{1j}^2; \tag{14.43}$$

and in the case $N < \infty$, to

$$\begin{aligned} &- N\frac{\ddot{\mathcal{D}}_\perp}{\mathcal{D}_\perp} = \mathrm{Ric}_{N,\nu}(\dot\gamma) \\ &+ \left\| U_\perp - \left(\frac{\mathrm{tr}\ U_\perp}{n-1}\right) I_{n-1} \right\|_{\mathrm{HS}}^2 + \frac{n-1}{(N-1)(N-n)} \left[\left(\frac{N-n}{n-1}\right) \mathrm{tr}\ U + \dot\gamma \cdot \nabla V(\gamma) \right]^2 + \sum_{j=2}^n u_{1j}^2; \end{aligned} \tag{14.44}$$

and as corollaries,

$$\begin{cases} \ddot\ell_\perp \geq \dfrac{(\dot\ell_\perp)^2}{N-1} + \mathrm{Ric}_{N,\nu}(\dot\gamma) \\ \\ -N\dfrac{\ddot{\mathcal{D}}_\perp}{\mathcal{D}_\perp} \geq \mathrm{Ric}_{N,\nu}(\dot\gamma). \end{cases} \tag{14.45}$$

Generalized Bochner formula and Γ_2 formalism

Of course there is an Eulerian translation of all that. This Eulerian formula can be derived either from the Lagrangian calculation, or from the Bochner formula, by a calculation parallel to the above one; the latter approach is conceptually simpler, while the former is faster. In any case the result is best expressed in terms of the differential operator

$$L = \Delta - \nabla V \cdot \nabla, \tag{14.46}$$

and can be written

$$\begin{aligned} L\,\frac{|\nabla\psi|^2}{2} - \nabla\psi\cdot\nabla L\psi &= \frac{(L\psi)^2}{N} + \mathrm{Ric}_{N,\nu}(\nabla\psi) \\ &+ \left(\left\| \nabla^2\psi - \left(\frac{\Delta\psi}{n}\right) I_n \right\|_{\mathrm{HS}}^2 + \frac{n}{N(N-n)} \left[\left(\frac{N-n}{n}\right) \Delta\psi + \nabla V\cdot\nabla\psi \right]^2 \right). \end{aligned} \tag{14.47}$$

It is convenient to reformulate this formula in terms of the Γ_2 formalism. Given a general linear operator L, one defines the associated Γ operator (or *carré du champ*) by the formula

$$\Gamma(f,g) = \frac{1}{2}\big[L(fg) - fLg - gLf\big].$$

Note that Γ is a bilinear operator, which in some sense encodes the deviation of L from being a derivation operator. In our case, for (14.46),

$$\Gamma(f,g) = \nabla f \cdot \nabla g.$$

Next introduce the Γ_2 operator (or *carré du champ itéré*) by

$$\Gamma_2(f,g) = \frac{1}{2}\big[L\Gamma(fg) - \Gamma(f,Lg) - \Gamma(g,Lf)\big].$$

In the case of (14.46), the important formula for later purpose is

$$\Gamma_2(\psi) := \Gamma_2(\psi,\psi) = L\frac{|\nabla\psi|^2}{2} - \nabla\psi\cdot\nabla(L\psi). \tag{14.48}$$

Then our previous computations can be rewritten as

$$\Gamma_2(\psi) = \frac{(L\psi)^2}{N} + \mathrm{Ric}_{N,\nu}(\nabla\psi) + \left(\left\| \nabla^2\psi - \left(\frac{\Delta\psi}{n}\right) I_n \right\|_{\mathrm{HS}}^2 + \frac{n}{N(N-n)} \left[\left(\frac{N-n}{n}\right) \Delta\psi + \nabla V \cdot \nabla\psi \right]^2 \right). \quad (14.49)$$

Of course, a trivial corollary is

$$\Gamma_2(\psi) \geq \frac{(L\psi)^2}{N} + \mathrm{Ric}_{N,\nu}(\nabla\psi). \quad (14.50)$$

And as the reader has certainly guessed, now one can take out the direction of motion (this computation is provided for completeness but will not be used): As before, define

$$\widehat{\nabla\psi} = \frac{\nabla\psi}{|\nabla\psi|},$$

then if f is a smooth function, let $\nabla^2_\perp f$ be $\nabla^2 f$ restricted to the space orthogonal to $\nabla\psi$, and $\Delta_\perp f = \mathrm{tr}\,(\nabla^2_\perp f)$, i.e.

$$\Delta_\perp f = \Delta f - \left\langle \nabla^2 f \cdot \widehat{\nabla\psi}, \widehat{\nabla\psi} \right\rangle,$$

and next,

$$L_\perp f = \Delta_\perp f - \nabla V \cdot \nabla f,$$

$$\Gamma_{2,\perp}(\psi) = L_\perp \frac{|\nabla\psi|^2}{2} - \nabla\psi \cdot \nabla(L_\perp\psi) - 2\left|(\nabla^2\psi)\cdot\widehat{\nabla\psi}\right|^2 - 2|(\nabla^2\psi)\cdot\widehat{\nabla\psi}|^2.$$

Then

$$\Gamma_{2,\perp}(\psi) = \frac{(L_\perp\psi)^2}{N-1} + \mathrm{Ric}_{N,\nu}(\nabla\psi) + \left\| \nabla^2_\perp\psi - \left(\frac{\Delta_\perp\psi}{n-1}\right) I_{n-1} \right\|^2 + \frac{n-1}{(N-1)(N-n)} \left[\left(\frac{N-n}{n-1}\right) \Delta_\perp\psi + \nabla V\cdot\nabla\psi \right]^2 + \sum_{j=2}^n (\partial_{1j}\psi)^2.$$

Curvature-Dimension bounds

It is convenient to declare that a Riemannian manifold M, equipped with its volume measure, satisfies the **curvature-dimension** estimate $\mathrm{CD}(K,N)$ if its Ricci curvature is bounded below by K and its dimension is bounded above by N: $\mathrm{Ric} \geq K$, $n \leq N$. (As usual, $\mathrm{Ric} \geq K$ is a shorthand for "$\forall x,\ \mathrm{Ric}_x \geq K g_x$.") The number K might be positive or negative. If the reference measure is not the volume, but $\nu = e^{-V}\mathrm{vol}$, then the correct definition is $\mathrm{Ric}_{N,\nu} \geq K$.

Most of the previous discussion is summarized by Theorem 17.15 below, which is all the reader needs to know about Ricci curvature to understand the rest of the proofs in these notes. For convenience I shall briefly recall the notation:

- measures: vol is the volume on M, $\nu = e^{-V}\,\mathrm{vol}$ is the reference measure;

- tensors: Ric is the Ricci curvature bilinear form, ∇^2 is the Hessian operator, $\mathrm{Ric}_{N,\nu}$ is the modified Ricci tensor defined by $\mathrm{Ric}_{N,\nu} = \mathrm{Ric} + \nabla^2 V - (\nabla V \otimes \nabla V)/(N-n)$, where the Hessian operator $\nabla^2 V(x)$ is identified with its associated bilinear form;

- operators: Δ is the Laplace(–Beltrami) operator on M, L is the modified Laplace operator defined by $L = \Delta - \nabla V \cdot \nabla$, and $\Gamma_2(\psi) = L(|\nabla\psi|^2/2) - \nabla\psi\cdot\nabla(L\psi)$;

- functions: ψ is an arbitrary function; in formulas involving the Γ_2 formalism it will be assumed to be of class C^3, while in formulas involving Jacobian determinants it will only be assumed to be semi-convex;

- geodesic paths: If ψ is a given function on M, $\gamma(t,x) = T_t(x) = \exp_x\big((t-t_0)\nabla\psi(x)\big)$ is the geodesic starting from x with velocity $\dot\gamma(t_0,x) = \nabla\psi(x)$, evaluated at time $t \in [0,1]$; it is assumed that $\mathcal{J}(t,x)$ does not vanish for $t \in (0,1)$; the starting time t_0 may be the origin $t_0 = 0$, but may also be any time in $[0,1]$;

- Jacobian determinants: $\mathcal{J}(t,x)$ is the Jacobian determinant of $T_t(x)$ (with respect to the reference measure ν, not with respect to the standard volume), $\ell = -\log \mathcal{J}$, and $\mathcal{D} = \mathcal{J}^{1/N}$ is the mean distortion associated with (T_t);

- the dot means differentiation with respect to time;

- finally, the subscript $\perp$ in $\mathcal{J}_\perp$, $\mathcal{D}_\perp$, $\Gamma_{2,\perp}$ means that the direction of motion $\dot\gamma = \nabla\psi$ has been taken out (see above for precise definitions).

Theorem 14.8. *Let M be a smooth Riemannian manifold of dimension n, and let $K \in \mathbb{R}$, $N \in [n,\infty]$. Then, the conditions below are all equivalent if they are required to hold true for arbitrary data; if they are fulfilled then M is said to satisfy a* $\mathrm{CD}(K,N)$ *curvature-dimension bound:*

(i) $\mathrm{Ric}_{N,\nu} \geq K$*;*

(ii) $\Gamma_2(\psi) \geq \dfrac{(L\psi)^2}{N} + K|\nabla\psi|^2$*;*

(iii) $\ddot\ell \geq \dfrac{(\dot\ell)^2}{N} + K|\dot\gamma|^2$.

If $N < \infty$, then this is also equivalent to

(iv) $\ddot{\mathcal{D}} + \left(\dfrac{K|\dot\gamma|^2}{N}\right)\mathcal{D} \leq 0$.

Moreover, these inequalities are also equivalent to

(ii') $\Gamma_{2,\perp}(\psi) \geq \dfrac{(L_\perp\psi)^2}{N-1} + K|\nabla\psi|^2$*;*

(iii') $\ddot\ell_\perp \geq \dfrac{(\dot\ell_\perp)^2}{N-1} + K|\dot\gamma|^2$*;*

and, in the case $N < \infty$,

(iv') $\ddot{\mathcal{D}}_\perp + \left(\dfrac{K|\dot\gamma|^2}{N-1}\right)\mathcal{D}_\perp \leq 0$.

Remark 14.9. Note carefully that the inequalities (i)-(iv') are required to be true *always*: For instance (ii) should be true for all ψ, all x and all $t \in (0,1)$. The equivalence is that [(i) true for all x] is equivalent to [(ii) true for all ψ, all x and all t], etc.

Examples 14.10 (One-dimensional $\mathrm{CD}(K,N)$ **model spaces).** (a) Let $K > 0$ and $1 < N < \infty$, consider

$$M = \left(-\sqrt{\frac{N-1}{K}}\,\frac{\pi}{2},\ \sqrt{\frac{N-1}{K}}\,\frac{\pi}{2}\right) \subset \mathbb{R},$$

equipped with the usual distance on $\mathbb{R}$, and the reference measure

$$\nu(dx) = \cos^{N-1}\left(\sqrt{\frac{K}{N-1}}\,x\right)\,dx;$$

then M satisfies $\mathrm{CD}(K,N)$, although the Hausdorff dimension of M is of course 1. Note that M is not complete, but this is not a problem since $\mathrm{CD}(K,N)$ is a local property. (We can also replace M by its closure, but then it is a manifold with boundary.)

(b) For $K<0$, $1 \le N < \infty$, the same conclusion holds true if one considers $M = \mathbb{R}$ and

$$\nu(dx) = \cosh^{N-1}\left(\sqrt{\frac{|K|}{N-1}}\, x\right) dx.$$

(c) For any $N \in [1,\infty)$, an example of one-dimensional space satisfying $\mathrm{CD}(0,N)$ is provided by $M = (0,+\infty)$, equipped with the reference measure $x^{N-1}\, dx$;

(d) For any $K \in \mathbb{R}$, take $M = \mathbb{R}$ and equip it with the reference measure

$$\nu(dx) = e^{-\frac{Kx^2}{2}}\, dx;$$

then M satisfies $\mathrm{CD}(K,\infty)$.

Sketch of proof of Theorem 14.8. It is clear from our discussion in this chapter that (i) implies (ii) and (iii); and (iii) is equivalent to (iv) by elementary manipulations about derivatives. (Moreover, (ii) and (iii) are equivalent modulo smoothness issues, by Eulerian/Lagrangian duality.)

It is less clear why, say, (ii) would imply (i). This comes from formulas (14.37) and (14.49). Indeed, assume (ii) and choose an arbitrary $x_0 \in M$, and $v_0 \in T_{x_0} M$. Construct a C^3 function ψ such that

$$\nabla\psi(x_0) = v_0, \qquad \nabla^2\psi(x_0) = \lambda_0\, I_n, \qquad \Delta\psi(x_0) (= n\lambda_0) = -\frac{n}{N-n}\big(\nabla V(x_0)\cdot v_0\big).$$

(This is fairly easy by using local coordinates, or distance and exponential functions.) Then all the remainder terms in (14.49) will vanish at x_0, so that

$$K|v_0|^2 = K|\nabla\psi(x_0)|^2 \le \left(\Gamma_2(\psi) - \frac{(L\psi)^2}{N}\right)(x_0) = \mathrm{Ric}_{N,\nu}\big(\nabla\psi(x_0)\big) = \mathrm{Ric}_{N,\nu}(v_0).$$

So indeed $\mathrm{Ric}_{N,\nu} \ge K$.

The proof goes in the same way for the equivalence between (i) and (ii'), (iii'), (iv'): again the problem is to understand why (ii') implies (i), and the reasoning is almost the same as before; the key point being that the extra error terms in $\partial_{1j}\psi$, $j \ne 2$, all vanish at x_0. □

Many interesting inequalities can be derived from $\mathrm{CD}(K,N)$. It was successfully advocated by Bakry and other authors during the past two decades that $\mathrm{CD}(K,N)$ should be considered as a property of the generalized Laplace operator L. On the contrary, it will be advocated in this course that $\mathrm{CD}(K,N)$ is a property of the solution of the optimal transport problem, when the cost function is the square of the geodesic distance. Of course, both points of view have their advantages and their drawbacks.

From differential to integral curvature-Dimension bounds

There are two ways to characterize the concavity of a function $f(t)$ on a time-interval, say $[0,1]$: the differential inequality $\ddot{f} \le 0$, or the integral bound $f\big((1-\lambda)\, t_0 + \lambda\, t_1\big) \ge$

$(1-\lambda)f(t_0)+\lambda f(t_1)$. If the latter is required to hold true for all $t_0, t_1 \in [0,1]$ and $\lambda \in [0,1]$, then the two formulations are equivalent.

There are two classical generalizations. The first one states that the differential inequality $\ddot{f} + K \le 0$ is equivalent to the integral inequality

$$f\big((1-\lambda)\,t_0 + \lambda\,t_1\big) \ge (1-\lambda)\,f(t_0) + \lambda\,f(t_1) + \frac{K\,t(1-t)}{2}\,(t_0-t_1)^2.$$

Another one is as follows: The differential inequality

$$\ddot{f}(t) + \Lambda f(t) \le 0 \tag{14.51}$$

is equivalent to the integral bound

$$f\Big((1-\lambda)\,t_0 + \lambda\,t_1\Big) \ge \tau^{(1-\lambda)}(|t_0-t_1|)\,f(t_0) + \tau^{(\lambda)}(|t_0-t_1|)\,f(t_1), \tag{14.52}$$

where

$$\tau^{(\lambda)}(\theta) = \begin{cases} \dfrac{\sin(\lambda\theta\sqrt{\Lambda})}{\sin(\theta\sqrt{\Lambda})} & \text{if } \Lambda > 0 \\ \lambda & \text{if } \Lambda = 0 \\ \dfrac{\sinh(\lambda\theta\sqrt{-\Lambda})}{\sinh(\theta\sqrt{-\Lambda})} & \text{if } \Lambda < 0. \end{cases}$$

A more precise statement and a proof are provided in a second appendix.

This leads to the following integral characterization of $\mathrm{CD}(K,N)$:

Theorem 14.11 (Integral curvature-dimension bounds). *Let M be a smooth Riemannian manifold, equipped with a reference measure $\nu = e^{-V}\,\mathrm{vol}$, and let d be the geodesic distance on M. Let $K \in \mathbb{R}$ and $N \in [1,\infty]$. Then, with the same notation as in Theorem 14.8, M satisfies $\mathrm{CD}(K,N)$, if and only if the following inequality is always true (for any semi-convex ψ, and almost any x, as soon as $\mathcal{J}(t,x)$ does not vanish for $t \in (0,1)$):*

$$\mathcal{D}(t,x) \ge \tau_{K,N}^{(1-t)}\,\mathcal{D}(0,x) + \tau_{K,N}^{(t)}\,\mathcal{D}(1,x) \qquad (N<\infty) \tag{14.53}$$

$$\ell(t,x) \le (1-t)\,\ell(0,x) + t\,\ell(1,x) - \frac{K\,t(1-t)}{2}\,d(x,y)^2 \qquad (N=\infty), \tag{14.54}$$

where $y = \exp_x(\nabla\psi(x))$ and, in case $N<\infty$,

$$\tau_{K,N}^{(t)} = \begin{cases} \dfrac{\sin(t\alpha)}{\sin\alpha} & \text{if } K>0 \\ t & \text{if } K=0 \\ \dfrac{\sinh(t\alpha)}{\sinh\alpha} & \text{if } K<0 \end{cases}$$

and

$$\alpha = \sqrt{\frac{|K|}{N}}\,d(x,y) \qquad (\alpha \in [0,\pi] \text{ if } K>0).$$

Proof of Theorem 14.11. If $N < \infty$, inequality (14.53) is obtained by transforming the differential bound of (iii) in Theorem 14.8 into an integral bound, after noticing that $|\dot\gamma|$ is a constant all along the geodesic γ, and equals $d(\gamma_0, \gamma_1)$. Conversely, to go from (14.53) to Theorem 14.8(iii), we select a geodesic γ, then reparametrize the geodesic $(\gamma_t)_{t_0 \le t \le t_1}$ into a geodesic on $[0,1]$, apply (14.53) to the reparametrized path and discover that

$$\mathcal{D}(t,x) \ge \tau_{K,N}^{(1-\lambda)}\, \mathcal{D}(t_0,x) + \tau_{K,N}^{(\lambda)}\, \mathcal{D}(t_1,x) \quad t = (1-\lambda)t_0 + \lambda t_1;$$

where now $\alpha = \sqrt{|K|/N}\, d(\gamma(t_0), \gamma(t_1))$. It follows that $\mathcal{D}(t,x)$ satisfies (14.52) for any choice of t_0, t_1; and this is equivalent to (14.51).

The reasoning is the same for the case $N = \infty$, starting from inequality (ii) in Theorem 14.8. □

The next result states that the the coefficients $\tau_{K,N}^{(t)}$ obtained in Theorem 14.11 can be *automatically improved* if N is finite and $K \neq 0$, by taking out the direction of motion:

Theorem 14.12 (Integral curvature-dimension bounds with direction of motion taken out). *Let M be a smooth Riemannian manifold, equipped with a reference measure $\nu = e^{-V}\mathrm{vol}$, and let d be the geodesic distance on M. Let $K \in \mathbb{R}$ and $N \in [1, \infty)$. Then, with the same notation as in Theorem 14.8, M satisfies* $\mathrm{CD}(K,N)$ *if and only if the following inequality is always true (for any semi-convex ψ, and almost any x, as soon as $\mathcal{J}(t,x)$ does not vanish for $t \in (0,1)$):*

$$\mathcal{D}(t,x) \ge \tau_{K,N}^{(1-t)}\, \mathcal{D}(0,x) + \tau_{K,N}^{(t)}\, \mathcal{D}(1,x) \tag{14.55}$$

where now

$$\tau_{K,N}^{(t)} = \begin{cases} t^{\frac{1}{N}} \left(\dfrac{\sin(t\alpha)}{\sin\alpha} \right)^{1-\frac{1}{N}} & \text{if } K > 0 \\ t & \text{if } K = 0 \\ t^{\frac{1}{N}} \left(\dfrac{\sinh(t\alpha)}{\sinh\alpha} \right)^{1-\frac{1}{N}} & \text{if } K < 0 \end{cases}$$

and

$$\alpha = \sqrt{\frac{|K|}{N-1}}\, d(x,y) \qquad (\alpha \in [0,\pi] \text{ if } K > 0).$$

Remark 14.13. When $N < \infty$ and $K > 0$ Theorem 14.12 contains the Bonnet–Myers theorem according to which $d(x,y) \le \pi\sqrt{(N-1)/K}$. With Theorem 14.11 the bound was only $\pi\sqrt{N/K}$.

Proof of Theorem 14.12. The proof that (14.55) implies $\mathrm{CD}(K,N)$ is done in the same way as for (14.53). (In fact (14.55) is stronger than (14.53).)

As for the other implication: Start from (14.22), and transform it into an integral bound:

$$\mathcal{D}_\perp(t,x) \ge \sigma_{K,N}^{(1-t)}\, \mathcal{D}_\perp(0,x) + \sigma_{K,N}^{(t)}\, \mathcal{D}_\perp(1,x),$$

where $\sigma_{K,N}^{(t)} = \sin(t\alpha)/\sin\alpha$ if $K > 0$; t if $K = 0$; $\sinh(t\alpha)/\sinh\alpha$ if $K < 0$. Next transform (14.19) into the integral bound

$$\mathcal{D}_{/\!/}(t,x) \geq (1-t)\,\mathcal{D}_{/\!/}(0,x) + t\,\mathcal{D}_{/\!/}(1,x).$$

Then both estimates can be combined thanks to Hölder's identity:

$$\begin{aligned}\mathcal{D}(t,x) &= \mathcal{D}_\perp(t,x)^{1-\frac{1}{N}}\,\mathcal{D}_{/\!/}(t,x)^{\frac{1}{N}} \\ &\geq \Big(\sigma_{K,N}^{(1-t)}\,\mathcal{D}(0,x) \,+\, \sigma_{K,N}^{(t)}\,\mathcal{D}(1,x)\Big)^{1-\frac{1}{N}} \Big((1-t)\,\mathcal{D}_{/\!/}(0,x) \,+\, t\,\mathcal{D}_{/\!/}(1,x)\Big)^{\frac{1}{N}} \\ &\geq (\sigma_{K,N}^{(1-t)})^{1-\frac{1}{N}}(1-t)^{\frac{1}{N}}\,\mathcal{D}(0,x) \,+\, (\sigma_{K,N}^{(t)})^{\frac{1}{N}} t^{\frac{1}{N}}\,\mathcal{D}_{/\!/}(1,x).\end{aligned}$$

Then inequality (14.55) follows. □

Estimate (14.55) is sharp in general. The following reformulation yields an appealing interpretation of $\mathrm{CD}(K,N)$ in terms of comparison spaces. In the sequel, I will write Jac_x for the (unoriented) Jacobian determinant evaluated at point x, computed with respect to a given reference measure.

Corollary 14.14 (curvature-dimension bounds by comparison). *Let M be a Riemannian manifold equipped with a reference measure $\nu = e^{-V}\mathrm{vol}$, $V \in C^2(M)$. Define the $\mathcal{J}$-function of M on $[0,1]\times\mathbb{R}_+\times\mathbb{R}_+$ by the formula*

$$\mathcal{J}_{M,\nu}(t,\delta,J) := \inf\Big\{\mathrm{Jac}_x(\exp(t\xi)); \quad |\xi(x)| = \delta; \quad \mathrm{Jac}_x(\exp(\xi)) = J\Big\}, \tag{14.56}$$

where the infimum is over all vector fields ξ defined around x, such that $\nabla\xi(x)$ is symmetric, and $\mathrm{Jac}_x(\exp_s\xi) \neq 0$ for $0 \leq s < 1$. Then, for any $K \in \mathbb{R}$, $N \in [1,\infty]$ ($K \leq 0$ if $N = 1$),

$$(M,\nu) \text{ satisfies } \mathrm{CD}(K,N) \Longleftrightarrow \mathcal{J}_{M,\nu} \geq \mathcal{J}^{(K,N)},$$

where $\mathcal{J}^{(K,N)}$ is the $\mathcal{J}$-function of the model space considered in Examples 14.10.

If N is an integer, then $\mathcal{J}^{(K,N)}$ is also the $\mathcal{J}$-function of the N-dimensional model space

$$S^{(K,N)} = \begin{cases} S^N\Big(\sqrt{\dfrac{N-1}{K}}\Big) & \text{if } K > 0, \\ \mathbb{R}^N & \text{if } K = 0, \\ \mathbb{H}^N\Big(\sqrt{\dfrac{N-1}{-K}}\Big) & \text{if } K < 0, \end{cases}$$

equipped with its volume measure.

Corollary 14.14 follows from Theorem 14.12 by a direct computation of the $\mathcal{J}$-function of the model spaces. In the case of $S^{(K,N)}$, one can also make a direct computation, or note that all the inequalities which we used to obtain (14.55) turn into equalities for suitable choices of parameters.

Remark 14.15. There is a quite similar (and more well-known) formulation of lower *sectional* curvature bounds which goes as follows. Define the $\mathcal{L}$-function of a manifold M by the formula

$$\mathcal{L}_M(t,\delta,L) = \inf\Big\{d\big(\exp_x(tv),\exp_x(tw)\big); \quad |v| = |w| = \delta;\ d(\exp_x v,\exp_x w) = L\Big\},$$

where the infimum is over tangent vectors $v, w \in T_x M$. Then M has sectional curvature larger than κ if and only if $\mathcal{L}_M \geq \mathcal{L}^{(\kappa)}$, where $\mathcal{L}^{(\kappa)}$ is the $\mathcal{L}$-function of the reference space $S^{(\kappa)}$, which is $S^2(1/\sqrt{\kappa})$ if $\kappa > 0$, $\mathbb{R}^2$ if $\kappa = 0$, and $\mathbb{H}^2(1/\sqrt{-\kappa})$ if $\kappa < 0$. By changing the infimum into a supremum, and by reversing the inequalities, one can also obtain a characterization of upper sectional curvature bounds. The comparison with (14.14) conveys the idea that sectional curvature bounds measure the rate of separation of geodesics in terms of distances, while Ricci curvature bounds do it in terms of Jacobian determinants.

Distortion coefficients

Apart from Definition 14.19, the material in this section is not necessary to the understanding of the rest of this course. Still, it is interesting because it will give a new interpretation of Ricci curvature bounds, and motivate the introduction of distortion coefficients, which will play a crucial role in the sequel.

Definition 14.16 (barycenters). *If A and B are two measurable sets in a Riemannian manifold, and $t \in [0,1]$, a t-barycenters of A and B is a point that can be written γ_t, where γ is a (minimizing, constant-speed) geodesic with $\gamma_0 = x$ and $\gamma_1 \in B_r(y)$. The set of all t-barycenters between A and B is denoted by $[A, B]_t$.*

Definition 14.17 (Distortion coefficients). *Let M be a Riemannian manifold, equipped with a reference measure $e^{-V}\mathrm{vol}$, $V \in C(M)$, and let x and y be any two points in M. Then the distortion coefficient $\overline{\beta}_t(x,y)$ between x and y at time $t \in (0,1)$ is defined as follows:*

- If x and y are joined by a unique geodesic γ, then

$$\overline{\beta}_t(x,y) = \lim_{r \to 0} \frac{\nu\big[[x, B_r(y)]_t\big]}{\nu[B_{tr}(y)]} = \lim_{r \to 0} \frac{\nu\big[[x, B_r(y)]_t\big]}{t^n\, \nu[B_r(y)]}; \tag{14.57}$$

- If x and y are joined by several minimizing geodesics, then

$$\overline{\beta}_t(x,y) = \inf_{\gamma} \limsup_{s \to 1^-} \overline{\beta}_t(x, \gamma_s), \tag{14.58}$$

where the infimum is over all minimizing geodesics joining x to y.
Finally, the values of $\overline{\beta}_t(x,y)$ for $t = 0$ and $t = 1$ are defined by

$$\overline{\beta}_1(x,y) \equiv 1; \qquad \overline{\beta}_0(x,y) := \liminf_{t \to 0^+} \overline{\beta}_t(x,y).$$

The heuristic meaning of distortion coefficients is as follows. Assume you are standing at point x and observing some device located at y. You are trying to estimate the volume of this device, but your appreciation is altered because light rays travel along curved lines (geodesics). If x and y are joined by a unique geodesic, then the coefficient $\overline{\beta}_0(x,y)$ tells by how much you are overestimating; so it is less than 1 in negative curvature, and greater than 1 in positive curvature. If x and y are joined by several geodesics, this is just the same, except that you choose to look in the direction where the device looks smallest.

More generally, $\overline{\beta}_t(x,y)$ compares the volume occupied by the light rays emanating from the light source, when they arrive close to $\gamma(t)$, to the volume that they would occupy in a flat space.

Now let us express distortion coefficients in differential terms, and more precisely *Jacobi fields*. A key concept in doing so will be the notion of *focal points*. The concept of focalization was already discussed in Chapter 8: A point y is said to be focal to another point

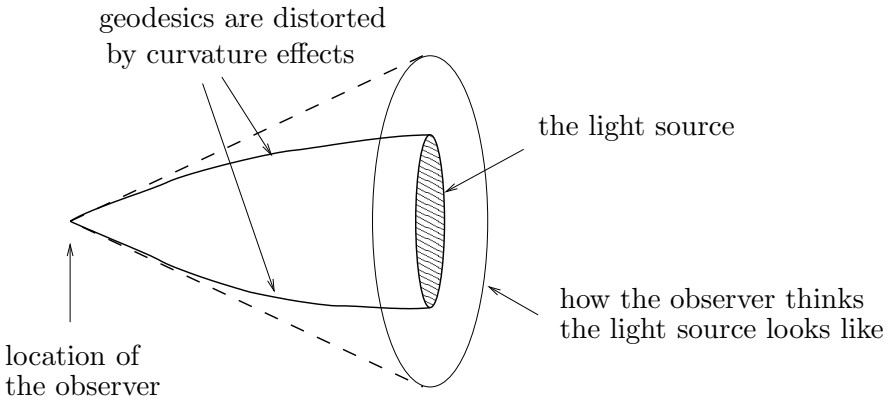


Fig. 14.4. Because of positive curvature effects, the observer overestimates the surface of the light source; in a negatively curved world this would be the contrary.

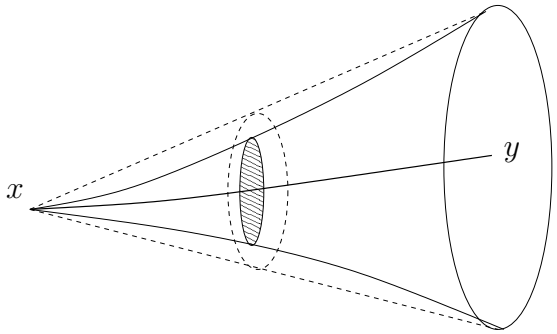


Fig. 14.5. The distortion coefficient is approximately equal to the ratio of the volume filled with lines, to the volume whose contour is in dashed line. Here the space is negatively curved and the distortion coefficient is less than 1.

x if there exists $v \in T_xM$ such that $y = \exp_v x$ and the differential $d_v \exp_x : T_xM \to T_yM$ is not invertible. It is equivalent to say that there is a geodesic γ which visits both x and y, and a Jacobi field J along γ such that $J(x) = 0$, $J(y) = 0$. This concept is obviously symmetric in x and y, and then x, y are said to be *conjugate points* (along γ).

If x and y are joined by a unique geodesic γ and are not conjugate, then by the local inversion theorem, for r small enough, there is a unique velocity $\xi(z)$ at $z \in B_r(y)$ such that $\exp_z \xi(z) = x$. Then the distortion coefficients can be interpreted as the Jacobian determinant of $\exp \xi$ at time t, renormalized by $(1-t)^n$, that would be the value in Euclidean space. The difference with the computations in the beginning of this chapter is that now the Jacobi field is not defined by its initial value and initial derivative, but rather by its initial value and its *final value*: $\exp_z \xi(z) = x$ independently of z, so the Jacobi field vanishes after a time 1. It will be convenient to reverse time so that $t = 0$ corresponds to x and $t = 1$ to y; so the conditions are $J(0) = 0$, $J(1) = I_n$. After that it is easy to derive the following

Proposition 14.18 (Computation of distortion coefficients). *Let M be a Riemannian manifold, let x and y be two points in M. Then*

$$\overline{\beta}_t(x, y) = \inf_\gamma \ \overline{\beta}_t^{[\gamma]}(x, y),$$

where the infimum is over all minimizing geodesics γ joining $\gamma(0) = x$ to $\gamma(1) = y$, and $\overline{\beta}_t^{[\gamma]}(x, y)$ is defined as follows:

- If x, y are not conjugate along γ, let $\mathbf{E}$ be an orthonormal basis of T_yM and define

$$\overline{\beta}_t^{[\gamma]}(x,y) = \begin{cases} \dfrac{\det \mathbf{J}^{0,1}(t)}{t^n} & \text{if } 0 < t \leq 1; \\ \\ \lim\limits_{s \to 0} \dfrac{\det \mathbf{J}^{0,1}(s)}{s^n} & \text{if } t = 0, \end{cases} \tag{14.59}$$

where $\mathbf{J}^{0,1}$ is the unique matrix of Jacobi fields along γ satisfying

$$\mathbf{J}^{0,1}(0) = 0; \qquad \mathbf{J}^{0,1}(1) = \mathbf{E};$$

- If x, y are conjugate along γ, define

$$\overline{\beta}_t^{[\gamma]}(x,y) = \begin{cases} 1 & \text{if } t = 1; \\ +\infty & \text{if } 0 \leq t < 1 \end{cases} .$$

Distortion coefficients can be explicitly computed for the model $\mathrm{CD}(K,N)$ spaces. These particular coefficients will play a key role in the sequel:

Definition 14.19 (Reference distortion coefficients). *Given $K \in \mathbb{R}$, $N \in [1, \infty]$ and $t \in [0,1]$, and two points x, y in some metric space $(\mathcal{X}, d)$, define $\beta_t^{(K,N)}(x,y)$ as follows:*

- If $0 < t \leq 1$ and $1 < N < \infty$ then

$$\beta_t^{(K,N)}(x,y) \;=\; \begin{cases} +\infty & \text{if } K > 0 \text{ and } \alpha > \pi, \\ \left(\frac{\sin(t\alpha)}{t \sin \alpha} \right)^{N-1} & \text{if } K > 0 \text{ and } \alpha \in [0, \pi], \\ 1 & \text{if } K = 0, \\ \left(\frac{\sinh(t\alpha)}{t \sinh \alpha} \right)^{N-1} & \text{if } K < 0, \end{cases} \tag{14.60}$$

where

$$\alpha \;=\; \sqrt{\frac{|K|}{N-1}}\, d(x,y). \tag{14.61}$$

- In the two limit cases $N \to 1$ and $N \to \infty$, modify the above expressions as follows:

$$\beta_t^{(K,1)}(x,y) \;=\; \begin{cases} +\infty & \text{if } K > 0, \\ 1 & \text{if } K \leq 0, \end{cases} \tag{14.62}$$

$$\beta_t^{(K,\infty)}(x,y) \;=\; e^{\frac{K}{6}(1-t^2)\, d(x,y)^2}. \tag{14.63}$$

- For $t = 0$ define $\beta_0^{(K,N)}(x,y) = 1$.

If $\mathcal{X}$ is the model space for $\mathrm{CD}(K,N)$, as in Examples 14.10, then $\beta^{(K,N)}$ is just the distortion coefficient on $\mathcal{X}$.

If K is positive then for fixed t, $\beta_t^{(K,N)}$ is an increasing function of α (going to $+\infty$ at $\alpha = \pi$); for fixed α, it is a decreasing function of t on $[0,1]$. All this is reversed for negative K. On the whole, $\beta_t^{(K,N)}$ is nondecreasing in K and nonincreasing in N.

The next two theorems relate distortion coefficients with the previous discussion about Ricci curvature lower bounds; they show that (a) distortion coefficients can be interpreted as the "best possible" coefficients in concavity estimates for the Jacobian determinant; (b) the curvature-dimension bound $\mathrm{CD}(K,N)$ is a particular case of a family of more general estimates characterized by a lower bound on the distortion coefficients.

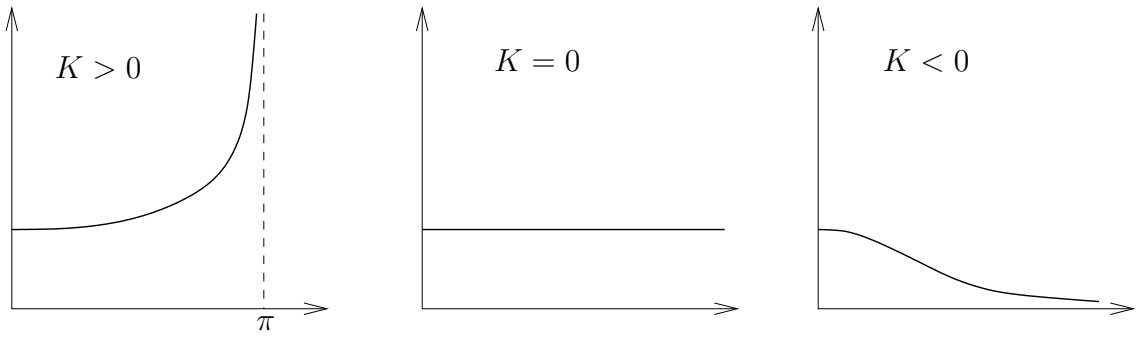


Fig. 14.6. The shape of the curves $\beta_t^{(K,N)}(x,y)$, as a function of $\alpha = (\sqrt{|K|}/(N-1))\, d(x,y)$.

Theorem 14.20 (Distortion coefficients and concavity of the Jacobian determinant). *Let M be a Riemannian manifold, and let x, y be any two points in M. Then if $(\beta_t(x,y))_{0\le t\le 1}$ and $(\beta_t(y,x))_{0\le t\le 1}$ are two families of nonnegative coefficients, the following two statements are equivalent:*

(a) $\forall t \in [0,1], \quad 0 \le \beta_t(x,y) \le \overline{\beta}_t(x,y); \quad \beta_t(y,x) \le \overline{\beta}_t(y,x)$;

(b) For any $N \ge n$, for any minimizing geodesic γ joining x to y, for any $t_0 \in [0,1]$, and for any initial vector field ξ around $x_0 = \gamma(t_0)$, $\nabla\xi(x_0)$ symmetric, let $\mathcal{J}(s)$ stand for the Jacobian determinant of $\exp((s-t_0)\xi)$ at x_0; if $\mathcal{J}(s)$ does not vanish for $0 < s < 1$, then for all $t \in [0,1]$,

$$\begin{cases} \mathcal{J}(t)^{\frac{1}{N}} \ge (1-t)\, \beta_{1-t}(y,x)^{\frac{1}{N}} \mathcal{J}(0)^{\frac{1}{N}} \;+\; t\, \beta_t(x,y)^{\frac{1}{N}} \mathcal{J}(1)^{\frac{1}{N}} \qquad (N < \infty) \\ \\ \log \mathcal{J}(0,x) \ge (1-t)\, \log \mathcal{J}(0) + t\, \log \mathcal{J}(1) \\ \qquad\qquad + \big[(1-t)\, \log \beta_{1-t}(y,x) + t\, \log \beta_t(x,y)\big] \qquad (N = \infty); \end{cases} \tag{14.64}$$

(c) Property (b) holds true for $N = n$.

Theorem 14.21 (Ricci curvature bounds in terms of distortion coefficients). *Let M be a Riemannian manifold with dimension n, equipped with its volume measure. Then the following two statements are equivalent:*

(a) $\mathrm{Ric} \ge K$;

(b) $\overline{\beta} \ge \beta^{(K,n)}$.

Sketch of proof of Theorem 14.20. To prove the implication (a)⇒(b), it suffices to establish (14.64) for $\beta = \overline{\beta}$. The case $N = \infty$ follows from the case $N < \infty$ by passing to the limit, since $\lim_{N\to 0}[N(a^{1/N} - 1)] = \log a$. So all we have to show is that if $n \le N < \infty$, then

$$\mathcal{J}(t)^{\frac{1}{N}} \ge (1-t)\, \overline{\beta}_{1-t}(y,x)^{\frac{1}{N}} \mathcal{J}(0)^{\frac{1}{N}} \;+\; t\, \overline{\beta}_t(x,y)^{\frac{1}{N}} \mathcal{J}(1)^{\frac{1}{N}}.$$

The case when x, y are conjugate can be treated by a limiting argument. (In fact the conclusion is that both $\mathcal{J}(0)$ and $\mathcal{J}(1)$ have to vanish if x and y are conjugate.) So we may assume that x and y are not conjugate, and then introduce a moving orthonormal basis $\mathbf{E}(t)$, along γ, and define the Jacobi matrices $\mathbf{J}^{1,0}(t)$ and $\mathbf{J}^{0,1}(t)$ by the requirement

$$J^{1,0}(0) = I_n, \quad J^{1,0}(1) = 0; \qquad J^{0,1}(0) = 0, \quad J^{0,1}(1) = I_n.$$

(Here $\mathbf{J}^{1,0}$ and $\mathbf{J}^{0,1}$ are identified with their expressions $J^{1,0}$ and $J^{0,1}$ in the moving basis $\mathbf{E}$.)

As noted after (14.7), the Jacobi equation is invariant under the change $t \to 1-t$, $\mathbf{E} \to -\mathbf{E}$, so $J^{1,0}$ becomes $J^{0,1}$ when one exchanges the roles of x and y, and replaces t by $1-t$. In particular, we have the formula

$$\frac{\det J^{0,1}(t)}{(1-t)^n} = \beta_{1-t}(y,x). \tag{14.65}$$

As in the beginning of this chapter, the issue is to compute the determinant at time t of a Jacobi field $J(t)$. Since the Jacobi fields are solutions of a linear differential equation of the form $\ddot{J} + RJ = 0$, they form a vector space of dimension $2n$, and they are invariant under right-multiplication by a constant matrix. It follows that

$$J(t) = J^{1,0}(t)\, J(0) \;+\; J^{0,1}(t)\, J(1). \tag{14.66}$$

The determinant in dimension n satisfies the following inequality: If X and Y are two $n \times n$ nonnegative symmetric matrices, then

$$\det(X+Y)^{\frac{1}{n}} \geq (\det X)^{\frac{1}{n}} + (\det Y)^{\frac{1}{n}}. \tag{14.67}$$

By combining this with the Hölder inequality, in the form

$$(a^{\frac{1}{n}} + b^{\frac{1}{n}})^{\frac{n}{N}} \geq (1-t)^{\frac{N-n}{n}} a^{\frac{1}{N}} + t^{\frac{N-n}{N}} b^{\frac{1}{N}},$$

we obtain the following generalization of (14.67):

$$\det(X+Y)^{\frac{1}{N}} \geq (1-t)^{\frac{N-n}{N}} (\det X)^{\frac{1}{N}} + t^{\frac{N-n}{N}} (\det Y)^{\frac{1}{N}}. \tag{14.68}$$

Then from (14.68) and (14.66) it results that

$$\begin{aligned}
(\det J(t))^{\frac{1}{N}} &\geq (1-t)^{\frac{N-n}{N}} (\det J^{1,0}(t))^{\frac{1}{N}} (\det J(0))^{\frac{1}{N}} + t^{\frac{N-n}{N}} (\det J^{0,1}(t))^{\frac{1}{N}} (\det J(1))^{\frac{1}{N}} \\
&= (1-t) \left[\frac{\det J^{1,0}(t)}{(1-t)^n}\right]^{\frac{1}{N}} \mathcal{J}(0)^{\frac{1}{N}} \;+\; t \left[\frac{\det J^{0,1}(t)}{t^n}\right]^{\frac{1}{N}} \mathcal{J}(1)^{\frac{1}{N}} \\
&= (1-t)\, \overline{\beta}_{1-t}(y,x)^{\frac{1}{N}}\, \mathcal{J}(0)^{\frac{1}{N}} \;+\; t\, \overline{\beta}_t(x,y)^{\frac{1}{N}}\, \mathcal{J}(1)^{\frac{1}{N}},
\end{aligned}$$

where (14.59) and (14.65) were used in the final step.

Next, it is obvious that (b) implies (c). To conclude the proof, it suffices to show that (c) $\Rightarrow$ (a). By symmetry and definition of $\overline{\beta}_t$, it is sufficient to show that $\beta_t(x,y) \leq \overline{\beta}_t^{[\gamma]}(x,y)$ for any geodesic γ. If x and y are conjugate along γ then there is nothing to prove. Otherwise, we can introduce $\xi(z)$ in the ball $B_r(y)$ such that for any $z \in B_r(y)$, $\exp_z \xi(z) = x$, and $\exp_z(t\xi(z))$ is the only geodesic joining z to x. Let then μ_0 be the uniform probability distribution on $B_r(y)$, and μ_1 be the Dirac mass at x; then $\exp \xi$ is the unique map T such that $T_\# \mu_0 = \mu_1$, so it is the optimal transport map, and therefore it can be written as $\exp(\nabla\psi)$ for some $d^2/2$-convex ψ; in particular, $\xi = \nabla\psi$. (Here I have chosen $t_0 = 1$, say.) So we can apply (b) with $N = n$, $\mathcal{D}(1) = 0$, $\mathcal{D}(0) = 1$, $\mathcal{D}(t,x) = \det J^{0,1}(t)$, and obtain

$$\det J^{0,1}(t) \geq t \beta_t(x,y)^{\frac{1}{n}}.$$

It follows that $\beta_t(x,y) \leq (\det J^{0,1}(t))/t^n = \overline{\beta}_t^{[\gamma]}(x,y)$, as desired. □

Sketch of proof of Theorem 14.21. To prove (a) $\Rightarrow$ (b), we apply inequality (14.55) with $n = N$, to conclude that Property (c) in Theorem 14.20 is satisfied with $\beta = \beta^{(K,n)}$; it follows that $\overline{\beta} \geq \beta^{(K,n)}$. Conversely, if $\overline{\beta} \geq \beta^{(K,n)}$, then Theorem 14.20 implies that inequality (14.55) is satisfied, which implies $\mathrm{CD}(K,n)$, or equivalently $\mathrm{Ric} \geq K$. □

Theorems 14.20 and 14.21 suggest a generalization of the $\mathrm{CD}(K,N)$ criterion: Given an effective dimension N, define the generalized distortion coefficients $\overline{\beta}_{N,\nu}$ as the best coefficients in (14.64) (the first inequality if $N < \infty$, the second one if $N = \infty$). Then the $\mathrm{CD}(K,N)$ inequality $\mathrm{Ric}_{N,\nu} \geq K$ is equivalent to the inequality $\overline{\beta}_{N,\nu} \geq \beta^{(K,N)}$. In this way we can see the condition $\mathrm{CD}(K,N)$ as a particular case of a more general condition $\mathrm{CD}(\beta,N)$, that would be defined by the inequality $\overline{\beta}_{N,\nu} \geq \beta$, where β would be, say, a given function of the distance between x and y. In the sequel of these notes, many results which hold true for $\mathrm{CD}(K,N)$ bounds would also hold true for this generalized curvature condition $\mathrm{CD}(\beta,N)$. The reasons why I shall not develop that more general theory are that (i) it is not clear at present that it really adds to the $\mathrm{CD}(K,N)$ theory; (ii) the condition $\mathrm{CD}(\beta,N)$ is in general nonlocal.

Remark 14.22. It is not a priori clear what kind of functions β can occur as distortion coefficients. It is striking to note that, in view of Theorems 14.12 and 14.11, for any given manifold M of dimension n the following two conditions are equivalent, say for $K > 0$:

$$\text{(i) } \forall x, y \in M,\ \forall t \in [0,1], \quad \overline{\beta}_t(x,y) \geq \left(\frac{\sin\Big(t \sqrt{\frac{K}{n}}\, d(x,y)\Big)}{t\, \sin\Big(\sqrt{\frac{K}{n}}\, d(x,y)\Big)} \right)^n ;$$

$$\text{(ii) } \forall x, y \in M,\ \forall t \in [0,1], \quad \overline{\beta}_t(x,y) \geq \left(\frac{\sin\Big(t \sqrt{\frac{K}{n-1}}\, d(x,y)\Big)}{t\, \sin\Big(\sqrt{\frac{K}{n-1}}\, d(x,y)\Big)} \right)^{n-1} .$$

This *self-improvement* property implies restrictions on the possible behavior of $\overline{\beta}$.

First Appendix: Second differentiability of convex functions

In this Appendix I shall provide a proof of Theorem 14.1. As explained right after the statement of that theorem, it suffices to consider the particular case of a convex function $\mathbb{R}^n \to \mathbb{R}$. So here is the statement to be proven:

Theorem 14.23 (Alexandrov's second differentiability theorem). *Let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then, for Lebesgue–almost every $x \in \mathbb{R}^n$, φ is differentiable at x and there exists a symmetric operator $A : \mathbb{R}^n \to \mathbb{R}^n$, characterized by any one of the following equivalent properties:*

(i) $\nabla\varphi(x+v) = \nabla\varphi(x) + Av + o(|v|)$ as $v \to 0$ (where v is such that φ is differentiable at $x+v$);

(i') $\partial\varphi(x+v) = \nabla\varphi(x) + Av + o(|v|)$ as $v \to 0$ (here $o(|v|)$ means a set whose elements are all bounded in norm like $o(|v|)$;

(ii) $\varphi(x+v) = \varphi(x) + \nabla\varphi(x) \cdot v + \dfrac{\langle Av, v\rangle}{2} + o(|v|^2)$ as $v \to 0$;

(ii') $\forall v \in \mathbb{R}^n, \qquad \varphi(x+tv) = \varphi(x) + t\nabla\varphi(x) \cdot v + t^2 \dfrac{\langle Av, v\rangle}{2} + o(t^2)$ as $t \to 0$.

The operator A is denoted by $\nabla^2\varphi(x)$ and called the Hessian of φ at x. When no confusion is possible, the quadratic form defined by A is also called the Hessian of φ at x. Moreover, the function $x \to \nabla^2\psi(x)$ (resp. $x \to \Delta\psi(x) = \mathrm{tr}\,(\nabla^2\psi(x))$ is the density of the absolutely continuous part of the distribution $\nabla^2_{\mathcal{D}'}\psi$ (resp. of the distribution $\Delta\psi$).

Before starting the proof, let me recall an elementary lemma about convex functions.

Lemma 14.24. *(i) Let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be a convex function, let $x_0, x_1, \dots, x_{n+1} \in \mathbb{R}^n$ such that $B(x_0, 2r)$ is included in the convex hull of $x_1, \dots, x_{n+1}$. Then,*

$$2\varphi(x_0) - \max_{1\le i\le n+1} \varphi(x_i) \le \inf_{B(x_0,2r)} \varphi \le \sup_{B(x_0,2r)} \varphi \le \max_{1\le i\le n+1} \varphi(x_i);$$

$$|\varphi\|_{\mathrm{Lip}(B(x_0,r))} \le \frac{2\Big(\max_{1\le i\le n+1} \varphi(x_i) - \varphi(x_0)\Big)}{r}.$$

(ii) If $(\varphi_k)_{k\in\mathbb{N}}$ is a sequence of convex functions which converges pointwise to some function Φ, then the convergence is locally uniform.

Proof of Lemma 14.24. If $x \in B(x_0, 2r)$ then of course $\varphi(x) \le \max(\varphi(x_1), \dots, \varphi(x_{n+1}))$. Next, if $z \in B(x_0, 2r)$, then $\widetilde{z} := 2x_0 - z \in B(x_0, 2r)$ and $\varphi(z) \ge 2\varphi(x_0) - \varphi(\widetilde{z}) \ge 2\varphi(x_0) - \max \varphi(x_i)$. Next, let $x \in B(x_0, r)$ and let $y \in \partial\varphi(x)$; let $z = x + ry/|y| \in B(x_0, r)$. Then from the subdifferential inequality, $r|y| = \langle y, z - x\rangle \le \varphi(z) - \varphi(x) \le 2(\max \varphi(x_i) - \varphi(x_0)$. This proves (i).

Let now $(\varphi_k)_{k\in\mathbb{N}}$ be a sequence of convex functions, let $x_0 \in \mathbb{R}^n$ and let $r > 0$. Let $x_1, \dots, x_{n+1}$ be such that $B(x_0, 2r)$ is included in the convex hull of $x_1, \dots, x_{n+1}$. If $\varphi_k(x_j)$ converges for all j, then by (i) there is a uniform bound on $\|\varphi_k\|_{\mathrm{Lip}}$ on $B(x_0, r)$. So if φ_k converges pointwise on $B(x_0, r)$, the convergence has to be uniform. This proves (ii). □

Now we start the proof of Theorem 14.23. To begin with, we should check that the formulations (i), (i'), (ii) and (ii') are equivalent.

Proof of the equivalence in Theorem 14.23. It is obvious that (i') ⇒ (i) and (ii) ⇒ (ii'), so we just have to show that (i) ⇒ (ii) and (ii') ⇒ (i').

To prove (i) ⇒ (ii), the idea is to use the mean value theorem; since a priori φ is not smooth, we shall regularize it. Let ζ be a radially symmetric nonnegative smooth function $\mathbb{R}^n \to \mathbb{R}$, with compact support in $B_1(0)$; and for any $\varepsilon > 0$ let $\zeta_\varepsilon(x) = \varepsilon^{-n}\zeta(x/\varepsilon)$; let then $\varphi_\varepsilon := \varphi * \zeta_\varepsilon$. The resulting function φ_ε is smooth and converges pointwise to φ as $\varepsilon \to 0$; moreover, since φ is locally Lipschitz we have (by dominated convergence) $\nabla\varphi_\varepsilon = (\nabla\varphi) * \zeta_\varepsilon$.

Then we can write

$$\begin{aligned}\varphi(x+v) - \varphi(x) &= \lim_{\varepsilon\to 0} \left[\varphi_\varepsilon(x+v) - \varphi_\varepsilon(x)\right] \\ &= \lim_{\varepsilon\to 0} \int_0^1 \nabla\varphi_\varepsilon(x+tv)\cdot v\, dt. \end{aligned} \tag{14.69}$$

Let us assume that $\varepsilon \le |v|$; then, by (i), for all $z \in B_{2\varepsilon}(x)$,

$$\nabla\varphi(z) = \nabla\varphi(x) + A(z-x) + o(|v|).$$

If $y \in B_\varepsilon(x)$, then we can integrate this identity against $\zeta_\varepsilon(y-z)\,dz$ (since $\zeta_\varepsilon(y-z) = 0$ for $|y-z| > \varepsilon$); taking into account $\int (z-x)\,\zeta_\varepsilon(z-x)\,dz = 0$, we obtain

$$\nabla\varphi_\varepsilon(y) = \nabla\varphi_\varepsilon(x) + A(y-x) + o(|v|).$$

In particular, $\nabla\varphi_\varepsilon(x+tv) = \nabla\varphi_\varepsilon(x) + tAv + o(|v|)$. By plugging this in the right-hand side of (14.69), we obtain Property (ii).

Now let us prove that (ii') ⇒ (i'). Without loss of generality we may assume that $x = 0$ and $\nabla\varphi(x) = 0$. So the assumption is $\varphi(tw) = t^2\,\langle Aw, w\rangle/2 + o(t^2)$, for any w. If (i') is false, then there are sequences $x_k \to 0$, $|x_k| \ne 0$, and $y_k \in \partial\varphi(x_k)$ such that

$$\frac{y_k - Ax_k}{|x_k|} \not\longrightarrow_{k\to\infty} 0. \tag{14.70}$$

Extract an arbitrary sequence from (x_k, y_k) (still denoted (x_k, y_k) for simplicity) and define

$$\varphi_k(w) := \frac{1}{|x_k|^2}\,\varphi(|x_k|w).$$

Assumption (ii) implies that φ_k converges pointwise to Φ defined by

$$\Phi(w) = \frac{\langle Aw, w\rangle}{2}.$$

The φ_k are convex, so the convergence is actually locally uniform by Lemma 14.24

Since $y_k \in \partial\varphi(x_k)$,

$$\forall z \in \mathbb{R}^n, \qquad \varphi(z) \geq \varphi(x_k) + \langle y_k, z - x_k\rangle,$$

or equivalently, with the notation $w_k = x_k/|x_k|$,

$$\forall w \in \mathbb{R}^n, \qquad \varphi_k(w) \geq \varphi_k(w_k) + \Big\langle \frac{y_k}{|x_k|}, w - w_k\Big\rangle. \tag{14.71}$$

The choice $w = w_k + y_k/|y_k|$ shows that $|y_k|/|x_k| \leq \varphi_k(w) - \varphi_k(w_k)$, so $|y_k|/|x_k|$ is bounded. Up to extraction of a subsequence, we may assume that $w_k = x_k/|x_k| \to \sigma$ and $y_k/|x_k| \to y$. Then we can pass to the limit in (14.71) and recover

$$\forall w \in \mathbb{R}^n, \qquad \Phi(w) \geq \Phi(\sigma) + \big\langle y, w - \sigma\big\rangle.$$

It follows that $y \in \partial\Phi(\sigma) = \{A\sigma\}$. So $y_k/|x_k| \to A\sigma$, or equivalently $(y_k - Ax_k)/|x_k| \to 0$. What we have shows is that each subsequence of our original sequence $(y_k - Ax_k)/|x_k|$ has a subsequence which converges to 0; it follows that the whole sequence converges to 0. This is in contradiction with (14.70), so (i') has to be true. □

Now, before proving Theorem 14.23 in full generality, we shall consider two particular cases which are much simpler.

Proof of Theorem 14.23 in dimension 1. Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a convex function. Then its derivative φ' is nondecreasing, and therefore differentiable almost everywhere. □

Proof of Theorem 14.23 when $\nabla\varphi$ is locally Lipschitz. Let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be a convex function, continuously differentiable with $\nabla\varphi$ locally Lipschitz. Then, by Rademacher's theorem, each function $\partial_i\varphi$ is differentiable almost everywhere, where ∂_i stands for the partial derivative with respect to x_i. So the functions $\partial_j(\partial_i\varphi)$ are defined almost everywhere. To conclude the proof, it suffices to show that $\partial_j(\partial_i\varphi) = \partial_i(\partial_j\varphi)$ almost everywhere. To prove this, let ζ be any C^2 compactly supported function; then, by successive use of the dominated convergence theorem and the smoothness of $\varphi * \zeta$,

$$(\partial_i\partial_j\varphi) * \zeta = \partial_i(\partial_j\varphi * \zeta) = \partial_i\partial_j(\varphi * \zeta) = \partial_j\partial_i(\varphi * \zeta) = \partial_j(\partial_i\varphi * \zeta) = (\partial_j\partial_j\varphi) * \zeta.$$

It follows that $(\partial_i\partial_j\varphi - \partial_j\partial_i\varphi) * \zeta = 0$, and since ζ is arbitrary this implies that $\partial_i\partial_j\varphi - \partial_j\partial_i\varphi$ vanishes almost everywhere. This concludes the argument. □

Proof of Theorem 14.23 in the general case. As in the proof of Theorem 10.8(ii), the strategy will be to reduce to the one-dimensional case. For any $v \in \mathbb{R}^n$, $t > 0$, and x such that φ is differentiable at x, define

$$Q_v(t,x) = \frac{\varphi(x+tv) - \varphi(x) - t\nabla\varphi(x)\cdot v}{t^2} \geq 0.$$

The goal is to show that for Lebesgue–almost all $x \in \mathbb{R}^n$,

$$q_v(x) := \lim_{t\to 0} Q_v(t,x)$$

exists for all v, and is a quadratic function of v.

Let Dom $q(x)$ be the set of $v \in \mathbb{R}^n$ such that $q_v(x)$ exists. It is clear from the definition that

(a) $q_v(x)$ is nonnegative and homogeneous of degree 2 in v on Dom $q(x)$;

(b) $q_v(x)$ is a convex function of v on Dom $q(x)$: this is just because it is the limit of the family $Q_v(t,x)$, which is convex in v;

(c) If v is interior to $\overline{\text{Dom } q(x)}$ and $q_w(x) \to \ell$ as $w \to v$, $w \in$ Dom $q(x)$, then also $v \in$ Dom $q(x)$ and $q_v(x) = \ell$. Indeed, let $\varepsilon > 0$ and let δ be so small that $|w - v| \leq \delta \Longrightarrow |q_w(x) - \ell| \leq \varepsilon$; then, we can find $v_1, \ldots, v_{n+1}$ in Dom $q(x) \cap B(v,\delta)$ so that v lies in the convex hull of $v_1, \ldots, v_{n+1}$, and then $v_0 \in$ Dom $q(x) \cap B(v,\delta)$ so that $v \in B(v_0,\delta)$ and $B(v_0,r)$ is included in the convex hull of $v_1, \ldots, v_{n+1}$. Then, by Lemma 14.24,

$$2Q_{v_0}(t,x) - \max Q_{v_i}(t,x) \leq Q_v(t,x) \leq \max Q_{v_i}(t,x).$$

Then,

$$\ell - 3\varepsilon \leq 2q_{v_0}(x) - \max q_{v_i}(x) \leq \liminf_{t\to 0} Q_v(t,x) \leq \limsup_{t\to 0} Q_v(t,x) \leq \max q_{v_i}(x) \leq \ell + \varepsilon.$$

It follows that $\lim Q_v(t,x) = \ell$, as desired.

Next, we use the same reasoning as in the proof of Rademacher's theorem (Theorem 10.8(ii)): Let v be given, $v \neq 0$, let us show that $q_v(x)$ exists for almost all x. By Fubini's theorem, it is sufficient to show that $q_v(x)$ exists λ_1–almost everywhere on each line parallel to v. So let $x_0 \in v^\perp$ be given, and let $L_{x_0} = x_0 + \mathbb{R}v$ be the line passing through x_0, parallel to v; the existence of $q_v(x_0 + t_0 v)$ is equivalent to the second differentiability of the convex function $\psi : t \to \varphi(x_0 + tv)$ at $t = t_0$, and from our study of the one–dimensional case we know that this happens for λ_1-almost all $t_0 \in \mathbb{R}$.

So we know that for each v, the set A_v of $x \in \mathbb{R}^n$ such that $q_v(x)$ does not exist is of zero measure. Let (v_k) be a dense subset of $\mathbb{R}^n$, and let $A = \cup A_{v_k}$: A is of zero measure, and for each $x \in \mathbb{R}^n \setminus A$, Dom $q(x)$ contains all the vectors v_k.

Let again $x \in \mathbb{R}^n \setminus A$. By Property (b), $q_v(x)$ is a convex function of v, so it is locally Lipschitz and can be extended uniquely into a continuous convex function $r(v)$ on $\mathbb{R}^n$. By Property (c), $r(v) = q_v(x)$, which means that Dom $q(x) = \mathbb{R}^n$.

At this point we know that for almost all x the limit $q_v(x)$ exists for all v, and it is a convex function of v, homogeneous of degree 2. What we do not know is whether $q_v(x)$ is a *quadratic* function of v.

Let us try to solve this problem by a regularization argument. Let ζ be a smooth nonnegative compactly supported function on $\mathbb{R}^n$, with $\int \zeta = 1$. Then $\nabla\varphi * \zeta = \nabla(\varphi * \zeta)$. Moreover, thanks to the nonnegativity of $Q_v(x,t)$ and Fatou's lemma,

$$\begin{aligned}(q_v * \zeta)(x) &= \int \lim_{t\downarrow 0} Q_v(y,t)\,\zeta(x-y)\,dy \\ &\le \liminf_{t\downarrow 0} \int Q_v(y,t)\,\zeta(x-y)\,dy \\ &= \liminf_{t\downarrow 0} \frac{1}{t^2}\Big[(\varphi*\zeta)(x+tv) - (\varphi*\zeta)(x) - t\,\nabla(\varphi*\zeta)(x)\cdot v\Big] \\ &= \frac{1}{2}\,\big\langle \nabla^2(\varphi*\zeta)(x)\cdot v,\, v\big\rangle.\end{aligned}$$

It is obvious that the right-hand side is a quadratic form in v, but this is only an *upper bound* on $q_v * \zeta(x)$. In fact, in general $q_v * \zeta \neq (1/2)\langle \nabla^2(\varphi*\zeta)v, v\rangle$. The difference is caused by the singular part of the measure $\mu_v := (1/2)\langle \nabla^2\varphi\cdot v, v\rangle$, defined in distribution sense by

$$\int \zeta(x)\,\mu_v(dx) = \frac{1}{2}\int \langle \nabla^2\zeta(x)\cdot v, v\rangle\,\varphi(x)\,dx.$$

This obstacle is the main new difficulty in the proof of Alexandrov's theorem, as compared to the proof of Rademacher's theorem.

To avoid the singular part of the measure μ_v, we shall appeal to Lebesgue's density theory, in the following precise form: Let μ be a locally finite measure on $\mathbb{R}^n$, and let $\rho\lambda_n + \mu_s$ be its Lebesgue decomposition into an absolutely continuous part and a singular part. Then, for Lebesgue–almost all $x \in \mathbb{R}^n$,

$$\frac{1}{\delta^n}\big\|\mu - \rho(x)\lambda_n\big\|_{\mathrm{TV}(B_\delta(x))} \xrightarrow[\delta\to 0]{} 0,$$

where $\|\cdot\|_{\mathrm{TV}(B_\delta(x))}$ stands for the total variation on the ball $B_\delta(x)$. Such an x will be called a Lebesgue point of μ.

So let ρ_v be the density of μ_v. It is easy to check that μ_v is locally finite, and we also showed that q_v is locally integrable. So, for λ_n–almost all x_0 we have

$$\frac{1}{\delta^n}\int_{B_\delta(x_0)} |q_v(x) - q_v(x_0)|\,dx \xrightarrow[\delta\to 0]{} 0; \qquad \frac{1}{\delta^n}\big\|\mu_v - \rho_v(x_0)\lambda_n\big\|_{\mathrm{TV}(B_\delta(x_0))} \xrightarrow[\delta\to 0]{} 0,$$

The goal is to show that $q_v(x_0) = \rho_v(x_0)$. Then the proof will be complete, since ρ_v is a quadratic form in v (indeed, $\rho_v(x_0)$ is obtained by averaging $\mu_v(dx)$, which itself is quadratic in v). Without loss of generality, we may assume that $x_0 = 0$.

To prove that $q_v(0) = \rho_v(0)$, it suffices to establish

$$\lim_{\delta\to 0}\frac{1}{\delta^n}\int_{B_\delta(0)} |q_v(x) - \rho_v(0)|\,dx = 0, \tag{14.72}$$

To estimate $q_v(x)$, we shall express it as a limit involving points in $B_\delta(x)$, and then use a Taylor formula; since φ is not a priori smooth, we shall go through a regularization procedure. Let ζ be as before, and let $\zeta_\varepsilon(x) = \varepsilon^{-n}\zeta(x/\varepsilon)$; let further $\varphi_\varepsilon := \varphi * \zeta$. We shall regularize φ on a scale $\varepsilon \le \delta$.

We can restrict the integral in (14.72) to those x such that $\nabla\varphi(x)$ exists and such that x is a Lebesgue point of $\nabla\varphi$; indeed, such points form a set of full measure. For such an x, $\varphi(x) = \lim_{\varepsilon\to 0}\varphi_\varepsilon(x)$, and $\nabla\varphi(x) = \lim_{\varepsilon\to 0}\nabla\varphi_\varepsilon(x)$. So,

$$\frac{1}{\delta^n}\int_{B_\delta(0)} |q_v(x) - \rho_v(0)|\,dx = \frac{1}{\delta^n}\int_{B_\delta(0)} \Big|\lim_{t\to 0}\Big[\frac{\varphi(x+t\delta v) - \varphi(x) - \nabla\varphi(x)\cdot t\delta v}{t^2\delta^2}\Big] - \rho_0(v)\Big|\,dx$$

$$
\begin{aligned}
&= \frac{1}{\delta^n} \int_{B_\delta(0)} \lim_{t\to 0}\ \lim_{\varepsilon\to 0} \Big| \frac{\varphi_\varepsilon(x+t\delta v) - \varphi_\varepsilon(x) - \nabla\varphi_\varepsilon(x)\cdot t\delta v}{t^2\delta^2} \ - \rho_0(v) \Big|\, dx \\
&= \frac{1}{\delta^n} \int_{B_\delta(0)} \lim_{t\to 0}\ \lim_{\varepsilon\to 0} \Big| \int_0^1 \Big[\langle \nabla^2\varphi_\varepsilon(x+st\delta v)\cdot v, v\rangle - 2\rho_v(0) \Big]\, (1-s)\, ds \Big|\, dx \\
&\le \liminf_{t\to 0}\ \liminf_{\varepsilon\to 0}\ \frac{1}{\delta^n} \int_{B_\delta(0)} \Big| \int_0^1 \big[\langle \nabla^2\varphi_\varepsilon(x+st\delta v)\cdot v, v\rangle - 2\rho_v(0) \big]\, (1-s)\, ds \Big|\, dx \\
&\le \liminf_{t\to 0}\ \liminf_{\varepsilon\to 0}\ \frac{1}{\delta^n} \int_0^1 \int_{B_\delta(st\delta v)} \big| \langle \nabla^2\varphi_\varepsilon(y)\cdot v, v\rangle - \rho_v(0) \big|\, dy\, ds,
\end{aligned}
$$

where Fatou's lemma and Fubini's theorem were used successively. Since $B(st\delta v, \delta) \subset B(0, (1+|v|)\delta)$, independently of s and t, we can bound the above expression by

$$
\begin{aligned}
&\liminf_{\varepsilon\to 0}\ \frac{1}{\delta^n} \int_{B(0,(1+|v|)\delta)} \big| \langle \nabla^2\varphi_\varepsilon(y)\cdot v, v\rangle - \rho_v(0) \big|\, dy \\
&\qquad = \liminf_{\varepsilon\to 0}\ \frac{1}{\delta^n} \int_{B(0,(1+|v|)\delta)} \Big| \int \zeta_\varepsilon(y-z)[\mu_v - \rho_v(0)\,\lambda_n](dz) \Big|\, dy \\
&\qquad \le \liminf_{\varepsilon\to 0}\ \frac{1}{\delta^n} \int_{B(0,(1+|v|)\delta)} \int \zeta_\varepsilon(y-z) |\mu_v - \rho_v(0)\,\lambda_n|(dz)\, dy.
\end{aligned}
$$

When y varies in $B(0,(1+|v|)\delta)$, z varies in $B(0,(1+|v|)\delta+\varepsilon) \subset B(0, C\delta)$ with $C = 2+|v|$. So, after using Fubini's theorem and integrating out $\zeta_\varepsilon(y-z)\, dy$, we conclude that

$$
\frac{1}{\delta^n} \int_{B_\delta(0)} |q_v(x) - \rho_v(0)|\, dx \le \|\mu_v - \rho_v(0)\,\lambda_n\|_{\mathrm{TV}(B(0,C\delta))}.
$$

The conclusion follows by taking the limit $\delta \to 0$.

Once $\nabla^2\varphi$ has been identified as the density of the distributional Hessian of φ, it follows immediately that $\Delta\varphi := \mathrm{tr}\,(\nabla^2\varphi)$ is the density of the distributional Laplacian of φ. (The trace of a matrix-valued nonnegative measure is singular if and only if the measure itself is singular.) □

Remark 14.25. The concept of distributional Hessian on a Riemannian manifold is a bit subtle, which is why I did not state anything about it in Theorem 14.1. On the other hand, there is no difficulty to define the distributional Laplacian.

Second Appendix: Very elementary comparison arguments

There are rather developed theories of comparison estimates for second-order linear differential equations; but the statement to be considered here can be proven by very elementary means.

Theorem 14.26. *Let $\Lambda \in \mathbb{R}$, and $f \in C([0,1]) \cap C^2(0,1)$, $f \ge 0$. Then the following two statements are equivalent:*

(i) $\ddot{f} + \Lambda f \le 0$ in $(0,1)$;

(ii) If $\Lambda < \pi^2$ then

$$
\forall t_0, t_1 \in [0,1] \qquad f\big((1-\lambda)t_0 + \lambda t_1\big) \ge \tau^{(1-\lambda)}(|t_0 - t_1|)\, f(t_0) \ + \ \tau^{(\lambda)}(|t_0 - t_1|)\, f(t_1),
$$

where

$$\tau^{(\lambda)}(t) = \begin{cases} \dfrac{\sin(\lambda\theta\sqrt{\Lambda})}{\sin(\theta\sqrt{\Lambda})} & \text{if } 0 < \Lambda < \pi^2 \\ \lambda & \text{if } \Lambda = 0 \\ \dfrac{\sinh(\lambda\theta\sqrt{-\Lambda})}{\sinh(\theta\sqrt{-\Lambda})} & \text{if } K < 0 \end{cases} .$$

If $\Lambda = \pi^2$ then $f(t) = c\sin(\pi t)$ for some $c \geq 0$; finally if $\Lambda > \pi^2$ then $f = 0$.

Proof of Theorem 14.26. The easy part is (ii) $\Rightarrow$ (i). If $\Lambda \geq \pi^2$ this is trivial. If $\Lambda < \pi^2$, take $\lambda = 1/2$, then a Taylor expansion shows that

$$\tau^{(1/2)}(\theta) = \frac{1}{2}\big(1 + \frac{\theta\Lambda^2}{8}\big) + o(\theta^3)$$

and

$$\frac{f(t_0) + f(t_1)}{2} = f\Big(\frac{t_0 + t_1}{2}\Big) + \frac{(t_0 - t_1)^2}{4}\ddot{f}\Big(\frac{t_0 + t_1}{2}\Big) + o(|t_0 - t_1|^2).$$

So, if we fix $t \in (0,1)$ and let $t_0, t_1 \to t$ in such a way that $t = (t_0 + t_1)/2$, we get

$$\tau^{(1/2)}(|t_0 - t_1|)\, f(t_0) + \tau^{(1/2)}(|t_0 - t_1|)\, f(t_1) - f(t) \;= \frac{(t_0 - t_1)^2}{8}\Big(\ddot{f}(t) + \Lambda f(t) + o(1)\Big).$$

By assumption the left-hand side is nonnegative, so in the limit we recover $\ddot{f} + \Lambda f \leq 0$.

Now consider the reverse implication (ii) $\Rightarrow$ (i). By abuse of notation, let us write $f(\lambda) = f((1-\lambda)t_0 + \lambda t_1)$, and define by a prime the derivation with respect to λ; so $f'' + \Lambda\theta^2 f \leq 0$, $\theta = |t_0 - t_1|$. Let $g(\lambda)$ be defined by the right-hand side of (ii); that is, $\lambda \to g(\lambda)$ is the solution of $g'' + \Lambda\theta^2 g = 0$ with $g(0) = f(0)$, $g(1) = f(1)$. The goal is to show that $f \geq g$ on $[0,1]$.

(a) Case $\Lambda < 0$. Let $a > 0$ be any constant; then $f_a := f + a$ still solves the same differential inequality as f, and $f_a > 0$ (even if we did not assume $f \geq 0$, we could take a large enough that this is true). Let g_a be defined as the solution of $g_a'' + \Lambda\theta^2 g_a = 0$ with $g_a(0) = f_a(0)$, $g_a(1) = f_a(1)$. As $a \to 0$, f_a converges to f and g_a converges to g, so it is sufficient to show $f_a \geq g_a$. Therefore, without loss of generality we may assume that f, g are positive, so g/f is continuous.

If g/f attains its maximum at 0 or 1, then we are done. Otherwise, there is $\lambda_0 \in (0,1)$ such that $(g/f)''(\lambda_0) \leq 0$, $(g/f)'(\lambda_0) = 0$, and then the identity

$$\left(\frac{g}{f}\right)'' = \frac{(g'' + \Lambda g)}{f} - \frac{g}{f^2}(f'' + \Lambda f) - 2\frac{f'}{f}\left(\frac{g}{f}\right)' - 2\Lambda\frac{g}{f},$$

evaluated at λ_0, yields $0 > -2\Lambda g/f$, which is impossible.

(b) Case $\Lambda = 0$. This is the basic property of concave functions.

(c) Case $0 < \Lambda < \pi^2$. Let $\theta = |t_0 - t_1| \leq 1$. Since $\theta\sqrt{\Lambda} < \pi$, we can find a function w such that $w'' + \Lambda\theta^2 w \leq 0$ and $w > 0$ on $(0,1)$. (Just take a well-chosen sine or cosine function.) Then $f_a := f + aw$ still satisfies the same differential inequality as f, and it is positive. Let g_a be defined by the equation $g_a'' + \Lambda\theta^2 g_a = 0$ with $g_a(0) = f_a(0)$, $g_a(1) = f_a(1)$. As $a \to 0$, $f_a \to f$ and $g_a \to g$, so it is sufficient to show that $f_a \geq g_a$. So we may assume that f and g are positive, and f/g is continuous.

Then the reasoning is parallel to the case $\Lambda < 0$: If f/g attains its minimum at 0 or 1, then we are done. Otherwise, there is $\lambda_0 \in (0,1)$ such that $(f/g)''(\lambda_0) \geq 0$, $(f/g)'(\lambda_0) = 0$, and then the identity

$$\left(\frac{f}{g}\right)'' = \frac{(f'' + \Lambda f)}{g} - \frac{f}{g^2}(g'' + \Lambda g) - 2\frac{g'}{g}\left(\frac{f}{g}\right)' - 2\Lambda\frac{f}{g},$$

evaluated at λ_0, yields $0 < -2\Lambda f/g$, which is impossible.

(d) Case $\Lambda = \pi^2$. Take $t_0 = 0$, $t_1 = 1$. Let then $g(\lambda) = \sin(\pi\lambda)$, and let $h := f/g$. The differential equations $f'' + \Lambda f \leq 0$ and $g'' + \Lambda g = 0$ combine to yield $(h'g^2)' = h''g^2 + 2gh'g' \leq 0$. So $h'g^2$ is nonincreasing. If $h'(\lambda_0) < 0$ for some $t_0 \in (0,1)$, then $h'g^2(\lambda_0) < 0$ for all $\lambda \geq \lambda_0$, so $h'(\lambda) \leq -C/(1-\lambda)^2$ as $\lambda \to 1$, where C is a positive constant. It follows that $h(\lambda)$ becomes negative for λ close to 1, which is impossible. If on the other hand $h'(\lambda_0) > 0$, then a similar reasoning shows that $h(\lambda)$ becomes negative for λ close to 0. The conclusion is that h' is identically 0, so f/g is a constant.

(e) If $\Lambda > \pi^2$, then for all $t_0, t_1 \in [0,1]$ with $|t_0 - t_1| = \pi/\sqrt{\Lambda}$, the function $f(\lambda) = f(\lambda t_0 + (1-\lambda t_1))$ is proportional to $\sin(\pi\lambda)$, by Case (d). By letting t_0, t_1 vary, it is easy to deduce that f is identically 0. □

Bibliographical Notes

Recommended textbooks about Riemannian geometry are the ones by do Carmo [142], Gallot, Hulin and Lafontaine [175] and Chavel [108]. All the necessary background about Hessians, Laplace–Beltrami operators, Jacobi fields and Jacobi equations can be found there. The discussion about the cut locus and regularity of the distance function is taken from [118].

Formula (14.1) does not seem to appear in standard textbooks of Riemannian geometry, but can be derived with the tools found therein, or by comparison with the sphere/hyperbolic space. On the sphere, the computation can be done directly, thanks to a classical formula of spherical trigonometry: If a, b, c are the lengths of the sides of a triangle drawn on the unit sphere S^2, and γ is the angle opposite to a, then $\cos c = \cos a \, \cos b + \sin a \, \sin b \, \cos\gamma$. A more standard computation usually found in textbooks is the asymptotic expansion of the perimeter of a circle centered at x with (geodesic) radius r, as $r \to 0$.

The differential inequalities relating the Jacobian of the exponential map with the Ricci curvature can be found (with minor variants) in a number of sources, e.g. [108, Section 3.4]. They usually appear in conjunction with volume comparison principles such as the Heintze–Kärcher, Lévy–Gromov, Bishop–Gromov (or Bishop–Gunther–Gromov) theorems, all of which express the idea that if the Ricci curvature is bounded below by K, and the dimension is less than N, then volumes along geodesic fields grow not faster than volumes in model spaces of constant sectional curvature having dimension N and Ricci curvature identicaly equal to K. These computations are usually performed in a smooth setting; their adaptation to the nonsmooth context of semi-convex functions has been achieved only recently, first by Cordero-Erausquin, McCann and Schmuckenschläger [118] (in a form that is somewhat different from the one presented here) and more recently by various sets of authors [117, 337, 247].

Bochner's formula appears e.g. as [175, Proposition 4.15] (for a vector field $\xi = \nabla\psi$) or as [296, Proposition 3.3 (3)] (for a vector field ξ such that $\nabla\xi$ is symmetric, i.e. the

1-form $p \to \xi \cdot p$ is closed). In both cases, it is derived from properties of the Riemannian curvature tensor. Another derivation of Bochner's formula for a gradient vector field is via the properties of the square distance function $d(x_0, x)^2$; this is quite simple, and not far from the presentation that I have followed, since $d(x_0, x)^2/2$ is the solution of the Hamilton–Jacobi equation at time 1, when the initial datum is 0 at x_0 and $+\infty$ everywhere else. But I thought that the explicit use of the Lagrangian/Eulerian duality would make Bochner's formula more intuitive to the readers (especially those who have some experience of fluid mechanics).

There are other Bochner formulas in the literature; Chapter 7 of Petersen's book [296] is entirely devoted to that subject. In fact "Bochner formula" is a generic name for many identities involving commutators of second-order differential operators and curvature.

The examples 14.10 are by now standard; they have been discussed for instance by Bakry and Qian [34], in relation with spectral gap estimates. When the dimension N is an integer, these reference spaces are of course obtained by projections of the model spaces with constant sectional curvature.

The practical importance of separating out the direction of motion is implicit in Cordero-Erausquin, McCann and Schmuckenschläger [118], but it was Sturm who attracted my attention on this. To implement this idea in the present chapter, I essentially followed the discussion in [339]. Also the integral bound (14.55) can be found in this reference.

Many analytic and geometric consequences of Ricci curvature bounds are discussed in Riemannian geometry textbooks such as the one by Gallot, Hulin and Lafontaine [175], and also in hundreds of research papers.

Cordero-Erausquin, McCann and Schmuckenschläger [118] express differential inequalities about the Jacobian determinant in terms of volume distortion coefficients; all the discussion about distortion coefficients is inspired from this reference.

About Bakry's approach to curvature-dimension bounds, among many sources one can consult the survey papers [30] and [236].

The almost everywhere second differentiability of convex functions was proven by Alexandrov in 1942 [8]. The proof which I gave in the first Appendix has several common points with the one that can be found in [156, pp. 241–245], but I have modified the argument to make it look as much as possible like the proof of Rademacher's theorem (Theorem 10.8(ii)). The resulting proof is a bit redundant in some respects, but hopefully it will look rather natural to the reader; also I think it is interesting to have a parallel presentation of the theorems by Rademacher and Alexandrov. Alberti and Ambrosio [4, Theorem 7.10] prove Alexandrov's theorem by a quite different technique, since they *deduce* it from Rademacher's theorem (in the form of the almost everywhere existence of the tangent plane to a Lipschitz graph) together with the area formula. Also they directly establish the differentiability of the gradient, and then deduce the existence of the Hessian; that is, they prove formulation (i) in Theorem 14.1 and then deduce (ii), while in the Appendix it was the other way round.

Lebesgue's density theorem can be found for instance in [156, p. 42]. The theorem according to which a nonincreasing function $\mathbb{R} \to \mathbb{R}$ is differentiable almost everywhere is a well-know result, that can be found as a corollary of [147, Theorems 7.2.4 and 7.2.7].

15

Otto calculus

Let M be a smooth complete Riemannian manifold, and let $P_2(M)$ be the associated Wasserstein space of order 2. Recall from Chapter 7 that $P_2(M)$ is a length space and that there is a nice representation formula for the Wasserstein distance W_2:

$$W_2(\mu_0, \mu_1)^2 = \inf \int_0^1 \|\dot{\mu}_t\|_{\mu_t}^2 \, dt, \tag{15.1}$$

where $\|\dot{\mu}\|_\mu$ is the norm of the infinitesimal variation $\dot{\mu}$ of the measure μ, defined by

$$\|\dot{\mu}\|_\mu = \inf \left\{ \int |v|^2 \, d\mu; \qquad \dot{\mu} + \nabla \cdot (v\mu) = 0 \right\}.$$

One of the reasons of the popularity of Riemannian geometry (as opposed to more general metric structures) is that they allow for rather explicit computations. At the end of the nineties, Otto realized that some precious help for intuition could be gained by performing computations of Riemannian nature in the Wasserstein space. His motivations will be described later on; to make a long story short, he needed a good formalism to study certain diffusive partial differential equations of which he knew that they could be considered as gradient flows in the Wasserstein space.

In this chapter, as in Otto's original papers, this problem will be considered from a purely formal point of view, and there will be no attempt of rigorous justification. So the problem is to set up rules for formally differentiating functions (i.e. functionals) on $P_2(M)$. To fix the ideas, and because this is an important example arising in many different contexts, I shall discuss only a certain class of functionals, that involve (i) a function $V : M \to \mathbb{R}$, used to distort the reference volume measure; and (ii) a function $U : \mathbb{R}_+ \to \mathbb{R}$, twice differentiable (at least on $(0, +\infty)$), which will relate the value of the density of our probability measure and the value of the functional. So let

$$\begin{cases} \nu(dx) := e^{-V(x)} \, \mathrm{vol}\,(dx) \\ \\ U_\nu(\mu) := \displaystyle\int_M U(\rho(x)) \, d\nu(x), \qquad \mu = \rho \, \nu. \end{cases} \tag{15.2}$$

So far the functional U_ν is only defined on the set of probability measures that are absolutely continuous with respect to ν, or equivalently with respect to the volume measure, and I shall not go beyond that setting before Part III of these notes. If ρ^0 stands for the density of μ with respect to the plain volume, then obviously $\rho^0 = e^{-V} \rho$, so there is the alternative expression

$$U_\nu(\mu) = \int_M U\big(e^V \rho^0\big) e^{-V}\, d\mathrm{vol}\,, \qquad \mu = \rho^0\, \mathrm{vol}\,.$$

One can think of U as a constitutive law for the **internal energy** of a fluid: this is jargon to say that the energy "contained" in a given fluid of density $\rho(x)$ is given by the formula $\int U(\rho)$. The function U should be a property of the fluid itself, and might reflect some microscopic interaction between particles of the fluid; it is natural to assume $U(0) = 0$.

In this thermodynamical analogy, one can also introduce the **pressure** law:

$$p(\rho) = \rho\, U'(\rho) - U(\rho). \tag{15.3}$$

The physical interpretation is as follows: if the fluid is enclosed in a domain Ω, then the pressure felt by the boundary $\partial\Omega$ at a point x is normal and proportional to $p(\rho)$ at that point. (Recall that the pressure is defined, up to a sign, as the partial derivative of the internal energy with respect to the volume of the fluid.) So if you consider a homogeneous fluid of total mass 1, in a volume V, then its density is $\rho = 1/V$, so the total energy is $V\, U(1/V)$, and the pressure should be $(-d/dV)[V\, U(1/V)] = p(1/V)$; this justifies formula (15.3).

To the pressure p is associated a total pressure $\int p(\rho)\, d\nu$, and one can again consider the influence of small variations of volume on this functional; this leads to the definition of the **iterated pressure**

$$p_2(\rho) = \rho p'(\rho) - p(\rho). \tag{15.4}$$

Both the pressure and the iterated pressure will appear naturally when one differentiates the energy functional: the pressure for first-order derivatives, and the iterated pressure for second-order derivatives.

Example 15.1. Let $m \neq 1$, and

$$U(\rho) = U^{(m)}(\rho) = \frac{\rho^m - \rho}{m-1};$$

then

$$p(\rho) = \rho^m, \qquad p_2(\rho) = (m-1)\, \rho^m.$$

There is an important limit case as $m \to 1$:

$$U^{(1)}(\rho) = \rho \log \rho;$$

then

$$p(\rho) = \rho, \qquad p_2(\rho) = 0.$$

By the way, the linear part $-\rho/(m-1)$ in $U^{(m)}$ does not contribute to the pressure, but has the merit to display the link between $U^{(m)}$ and $U^{(1)}$.

Differential operators will also be useful. Let Δ be the Laplace(–Beltrami) operator on M, then the distortion of the volume element by the function V leads to a natural second-order operator:

$$L = \Delta - \nabla V \cdot \nabla. \tag{15.5}$$

Recall from Chapter 14 the expression of the *carré du champ itéré* associated with L:

$$\Gamma_2(\psi) = L\Big(\frac{|\nabla\psi|^2}{2}\Big) - \nabla\psi \cdot \nabla(L\psi) \tag{15.6}$$

$$= \|\nabla^2\psi\|^2_{\mathrm{HS}} + \big(\mathrm{Ric} + \nabla^2 V\big)(\nabla\psi), \tag{15.7}$$

where the second equality is a consequence of Bochner's formula.

The next formula is the first important result in this chapter: it gives an "explicit" expression for the gradient of the functional U_ν. For a given measure μ, the gradient of U_ν at μ is a "tangent vector" at μ in the Wasserstein space, so this should be an infinitesimal variation of μ.

Formula 15.2 (gradient formula in Wasserstein space). *Let μ be absolutely continuous with respect to ν. Then, with the above notation,*

$$\mathrm{grad}_\mu U_\nu = -\nabla \cdot \big(\mu\, \nabla U'(\rho)\big) \tag{15.8}$$

$$= -\nabla \cdot \big(e^{-V} \nabla p(\rho)\big). \tag{15.9}$$

Remark 15.3. The first expression in the right-hand side of (15.8) is the divergence of a vector-valued measure; recall that $\nabla \cdot m$ is defined in weak sense by its action on compactly supported smooth functions:

$$\int \phi\, d(\nabla \cdot m) = -\int \nabla\phi \cdot (dm).$$

The second expression in the right-hand side of (15.8) is the divergence of a vector field; the volume measure is implicit, so I could have written, more rigorously,

$$\mathrm{grad}_\mu U_\nu = -\nabla \cdot \big(e^{-V} \nabla p(\rho)\big)\, \mathrm{vol}\,.$$

Both expressions in (15.8) are interesting, the first one because it writes the "tangent vector" $\mathrm{grad}_\mu U_\nu$ in the "normalized" form $-\nabla \cdot (\mu \nabla \psi)$, with $\psi = U'(\rho)$, and the second one gives because it gives the result as the divergence of a vector field.

Here below are some important examples of application of Formula 15.2.

Example 15.4. Assume $\nu = \mathrm{vol}$ and define the H-functional of Boltzmann (opposite of the entropy) by

$$H(\mu) = \int_M \rho \log \rho.$$

Then the second expression in equation (15.8) yields

$$\mathrm{grad}_\mu H = -\Delta \mu,$$

which can be identified with the function $-\Delta\rho$. So *the gradient of Boltzmann's entropy is the Laplace operator.* This short statement is one of the first striking conclusions of Otto's formalism.

Example 15.5. Now consider a general $\nu = e^{-V}\mathrm{vol}\,$, write $\mu = \rho\,\nu = \rho^0 \mathrm{vol}\,$, and define

$$H_\nu(\mu) = \int_M \rho \log \rho\, d\nu = \int_M (\log \rho^0 + V)\, d\mu$$

(this is the H-functional relative to the reference measure ν).

$$\mathrm{grad}_\mu H_\nu = -\left(\Delta\rho - \nabla V \cdot \nabla \rho\right) \nu = -\left(\Delta\rho - \nabla V \cdot \nabla\rho\right) e^{-V}\, \mathrm{vol}\,.$$

So this can be identified with the function $-e^{-V}(L\rho)$, where

$$L\rho = \Delta\rho - \nabla V \cdot \nabla \rho = e^V \nabla \cdot (e^{-V} \nabla \rho).$$

In short, *the gradient of the relative entropy is the distorted Laplace operator.*

Example 15.6. To generalize Example 15.4 in another direction, consider

$$H^{(m)}(\mu) = \int \frac{\rho^m - \rho}{m-1}\, d\mathrm{vol}\,.$$

Then

$$\mathrm{grad}_\mu H^{(m)} = -\Delta(\rho^m).$$

More generally, if ρ is the density with respect to an arbitrary reference measure ν, and

$$H_\nu^{(m)}(\mu) = \int \frac{\rho^m - \rho}{m-1}\, d\nu,$$

then

$$\mathrm{grad}_\mu U_\nu = -e^V \nabla \cdot \big(e^{-V} \nabla \rho^m\big). \tag{15.10}$$

The next formula is about second-order derivatives, or Hessians. Since the Hessian of U_ν at μ is a quadratic form on the tangent space $T_\mu P_2$, I shall write down its expression when evaluated on a tangent vector of the form $-\nabla \cdot (\mu \nabla \psi)$.

Formula 15.7 (Hessian formula in Wasserstein space). *Let μ be absolutely continuous with respect to ν, and let $\dot\mu = -\nabla \cdot (\mu \nabla \psi)$ be a tangent vector at μ. Then, with the above notation,*

$$\mathrm{Hess}_\mu U_\nu(\dot\mu) = \int_M \Gamma_2(\psi)\, p(\rho)\, d\nu + \int_M (L\psi)^2\, p_2(\rho)\, d\nu \tag{15.11}$$

$$= \int_M \Big[\|\nabla^2 \psi\|_{\mathrm{HS}}^2 + \big(\mathrm{Ric} + \nabla^2 V\big)(\nabla\psi)\Big] p(\rho) d\nu + \int_M \Big(-\Delta\psi + \nabla V \cdot \nabla \psi\Big)^2 p_2(\rho)\, d\nu. \tag{15.12}$$

Remark 15.8. As expected, this is a quadratic expression in $\nabla\psi$ and its derivatives; and this expression does depend on the measure μ.

Example 15.9. Applying the formula with $U(\rho) = (\rho^m - \rho)/(m-1)$, recalling that $\mu = \rho\nu$, one obtains

$$\mathrm{Hess}_\mu H_\nu^{(m)}(\dot\mu) = \int_M \Big(\|\nabla^2\psi\|_{\mathrm{HS}}^2 + (\mathrm{Ric} + \nabla^2 V)(\nabla\psi) + (m-1)\big(\Delta\psi - \nabla V \cdot \nabla\psi\big)^2\Big)\, \rho^{m-1}\, d\mu.$$

In the limit case $m = 1$, which is $U(\rho) = \rho \log \rho$, this expression simplifies into

$$\mathrm{Hess}_\mu H_\nu(\dot\mu) = \int_M \Big(\|\nabla^2\psi\|_{\mathrm{HS}}^2 + (\mathrm{Ric} + \nabla^2 V)(\nabla\psi)\Big)\, d\mu.$$

With the notation of Chapter 14, this can be rewritten

$$\mathrm{Hess}_\mu H_\nu(\dot\mu) = \int_M \Big(\|\nabla^2\psi\|_{\mathrm{HS}}^2 + \mathrm{Ric}_{\infty,\nu}(\nabla\psi)\Big)\, d\mu.$$

Formulas 15.2 and 15.7 will only be justified at a heuristic level. A rigorous proof would require much more definitions and apparatus, as well as regularity and decay assumptions on the measures and the functionals. So here I shall disregard all issues about integrability and regularity, which will be a huge simplification. Still, the proofs will not be completely trivial.

"Proof" of Formula 15.2. As usual, when the integration measure is not specified, it will be the volume. To understand the proof, it is important to make the distinction between a *gradient* and a *differential*.

Let ζ be such that the tangent vector $\mathrm{grad}_\mu U_\nu$ can be represented as $-\nabla\cdot(\mu\nabla\zeta)$, and let $\partial_t\mu = -\nabla\cdot(\mu\nabla\psi)$ be an arbitrary "tangent vector". The infinitesimal variation of the density $\rho = d\mu/d\nu$ is given by

$$\partial_t\rho = -e^V\nabla\cdot\big(\rho e^{-V}\nabla\psi\big).$$

By direct computation and integration by parts, the infinitesimal variation of U_ν along that variation is equal to

$$\begin{aligned}\int U'(\rho)\,\partial_t\rho\,d\nu &= -\int U'(\rho)\;\nabla\cdot(\rho\,e^{-V}\nabla\psi)\\ &= \int \nabla U'(\rho)\cdot\nabla\psi\;\rho\,e^{-V}\\ &= \int \nabla U'(\rho)\cdot\nabla\psi\;d\mu.\end{aligned}$$

By definition of the gradient operator, this should coincide with

$$\big\langle \mathrm{grad}_\mu U_\nu,\,\partial_t\mu\big\rangle = \int\nabla\zeta\cdot\nabla\psi\,d\mu.$$

If this should hold true for all ψ, the only possible choice is that $\nabla U'(\rho) = \nabla\zeta(\rho)$, at least μ-almost everywhere. In any case $\zeta := U'(\rho)$ provides an admissible representation of $\mathrm{grad}_\mu U_\nu$. This proves the formula for the gradient. □

For the second order (Formula (15.7)), things are more intricate. The following identity will be helpful: if ξ is a tangent vector at x on a Riemannian manifold $\mathcal{M}$, and F is a function on $\mathcal{M}$, then

$$\mathrm{Hess}_x F(\xi) = \left.\frac{d^2}{dt^2}\right|_{t=0} F(\gamma(t)), \tag{15.13}$$

where $\gamma(t)$ is a geodesic starting from $\gamma(0) = x$ with velocity $\dot\gamma(0) = \xi$. To prove (15.13), it suffices to note that the first derivative of $F(\gamma(t))$ is $\dot\gamma(t)\cdot\nabla F(\gamma(t))$; so the second derivative is $(d/dt)(\dot\gamma(t))\cdot\nabla F(\gamma(t)) + \big\langle\nabla^2 F(\gamma(t))\cdot\dot\gamma(t),\dot\gamma(t)\big\rangle$, and the first term vanishes because a geodesic has zero acceleration.

"Proof" of Formula 15.7. The problem consists in differentiating $U_\nu(\mu_t)$ twice along a geodesic path of the form

$$\begin{cases}\partial_t\mu + \nabla\cdot(\mu\nabla\psi) = 0\\[2ex] \partial_t\psi + \dfrac{|\nabla\psi|^2}{2} = 0.\end{cases}$$

The following integration by parts formula will be useful:

$$\int\nabla f\cdot\nabla g\,d\nu = -\int (Lf)g\,d\nu, \qquad L = \Delta - \nabla V\cdot\nabla. \tag{15.14}$$

From the proof of the gradient formula, one has, with the notation $\mu_t = \rho_t\nu$,

$$\frac{dU_\nu(\mu_t)}{dt} = \int_M \nabla\psi_t\cdot\nabla U'(\rho_t)\rho_t\,d\nu.$$

It follows from the definition of p that $p'(\rho) = \rho\, U''(\rho)$, and so

$$\nabla U'(\rho)\rho = \rho\, U''(\rho)\nabla\rho = p'(\rho)\nabla\rho = \nabla p(\rho).$$

So

$$\frac{dU_\nu(\mu_t)}{dt} = \int_M \nabla\psi_t \cdot \nabla p(\rho_t)\, d\nu \tag{15.15}$$

$$= -\int_M (L\psi_t)\, p(\rho_t)\, d\nu. \tag{15.16}$$

It remains to differentiate again. To alleviate notation, I shall not write explicitly the time variable. So

$$\frac{d^2U_\nu(\mu)}{dt^2} = -\int \big(L\partial_t\psi\big)p(\rho)\, d\nu - \int (L\psi)p'(\rho)\partial_t\rho\, d\nu \tag{15.17}$$

$$= \int L\left(\frac{|\nabla\psi|^2}{2}\right) p(\rho)\, d\nu - \int (L\psi)p'(\rho)\partial_t\mu. \tag{15.18}$$

The last term in (15.18) can be rewritten as

$$\int (L\psi)\, p'(\rho)\, \nabla\cdot(\mu\nabla\psi)$$

$$= -\int \nabla\big((L\psi)p'(\rho)\big)\cdot\nabla\psi\, d\mu$$

$$= -\int \nabla\big((L\psi)p'(\rho)\big)\cdot\nabla\psi\,\rho\, d\nu$$

$$= -\int \nabla(L\psi)\cdot\nabla\psi\, p'(\rho)\,\rho\, d\nu - \int (L\psi)p''(\rho)\,\rho\,\nabla\rho\cdot\nabla\psi\, d\nu$$

$$= -\int \nabla(L\psi)\cdot\nabla\psi\,\rho\, p'(\rho)\, d\nu - \int (L\psi)\nabla p_2(\rho)\cdot\nabla\psi\, d\nu. \tag{15.19}$$

The second term in (15.19) needs a bit of reworking: it can be recast as

$$-\int \nabla\big(L\psi\, p_2(\rho)\big)\cdot\nabla\psi\, d\nu - \int (\nabla L\psi)p_2(\rho)\cdot\nabla\psi\, d\nu$$

$$= \int (L\psi)^2 p_2(\rho)\, d\nu - \int (\nabla L\psi)\cdot\nabla\psi\, p_2(\rho)\, d\nu,$$

where formula (15.14) has been used once more.

By collecting all these calculations,

$$\frac{d^2U_\nu(\mu)}{dt^2} = \int L\left(\frac{|\nabla\psi|^2}{2}\right) p(\rho)\, d\nu + \int (L\psi)^2 p_2(\rho)\, d\nu$$
$$+ \int (\nabla\psi\cdot\nabla L\psi)\,\big(p_2(\rho) - \rho\, p'(\rho)\big)\, d\nu.$$

Since $p_2(\rho) - \rho\, p'(\rho) = -p(\rho)$, this transforms into

$$\int \left(L\left(\frac{|\nabla\psi|^2}{2}\right) - \nabla\psi\cdot\nabla L\psi\right) p(\rho)\, d\nu + \int (L\psi)^2 p_2(\rho). \tag{15.20}$$

In view of (15.6), this establishes formula (15.11).

To obtain formula (15.12), it is sufficient to prove the identity (15.7). In the case when $V = 0$, this is just the Bochner formula (14.28). When V is nonzero, there is an additional term:

$$\begin{aligned} &- \nabla V \cdot \nabla \frac{|\nabla \psi|^2}{2} + \nabla \psi \cdot \nabla (\nabla V \cdot \nabla \psi) \\ =& - \big\langle \nabla^2 \psi \cdot \nabla V, \nabla \psi \big\rangle + \big\langle \nabla^2 V \cdot \nabla \psi, \nabla \psi \big\rangle + \big\langle \nabla^2 \psi \cdot \nabla V, \nabla \psi \big\rangle \\ =& \big\langle \nabla^2 V \cdot \nabla \psi, \nabla \psi \big\rangle. \end{aligned}$$

This proves (15.7), and concludes the argument. □

Exercise 15.10. "Prove" that the gradient of an arbitrary functional $\mathcal{F}$, on $P_2(M)$ can be written

$$\mathrm{grad}_\mu \, \mathcal{F} = -\nabla \cdot (\mu \nabla \phi), \qquad \phi = \frac{\delta \mathcal{F}}{\delta \mu},$$

where $\delta \mathcal{F} / \delta \mu$ is a function defined by

$$\frac{d}{dt} \mathcal{F}(\mu_t) = \int \left(\frac{\delta \mathcal{F}}{\delta \mu} \right) \partial_t \mu_t.$$

Check that in the particular case

$$\mathcal{F}(\mu) = \int_M F\Big(x, \rho(x), \nabla \rho(x)\Big) \, d\nu(x), \tag{15.21}$$

where $F = F(x, \rho, p)$ is a smooth function of $\rho \in \mathbb{R}_+$, $(x, p) \in TM$, one has

$$\left(\frac{\partial \mathcal{F}}{\partial \mu} \right)(x) = (\partial_\rho F)\big(x, \rho(x), \nabla \rho(x)\big) - (\nabla_x - \nabla V(x)) \cdot (\nabla_p F)\big(x, \rho(x), \nabla \rho(x)\big)$$

The following two open problems (loosely formulated) are natural and interesting, and I don't know how difficult they are:

Open Problem 15.11. *Find a nice formula for the Hessian of the functional $\mathcal{F}$ appearing in* (15.21).

Open Problem 15.12. *Find a nice formalism playing the role of the Otto calculus in the space $P_p(M)$, for $p \neq 2$. More generally, are there nice formal rules for taking derivatives along displacement interpolation, for general Lagrangian cost functions?*

To conclude this chapter, I shall come back to the subject of rigorous justification of Otto's formalism. At the time of writing, several theories have been developed, at least in Euclidean setting (see the bibliographical notes); but they are rather heavy and not completely convincing (I can afford this negative comment since I myself participated in the story). From the technical point of view, they are based on the natural strategy which consists in truncating and regularizing, then apply the arguments presented in this chapter, then passing to the limit.

A quite different strategy, which I personally recommend, consists in translating all the Eulerian statements in the language of Lagrangian formalism. This is less appealing for intuition and calculations, but somehow easier to justify in the case of optimal transport. For instance, instead of the Hessian operator, one will only speak of the second derivative

along geodesics in the Wasserstein space. This point of view will be developed in the next two chapters, and then a rigorous treatment will not be that painful.

Still, in many situations the Eulerian point of view is better for intuition and for understanding, in particular in certain problems involving functional inequalities. The above discussion might be summarized by the slogan *"Think Eulerian, prove Lagrangian"*. This is a rather exceptional situation from the point of view of fluid dynamics, where the standard would rather be "Think Lagrangian, prove Eulerian" (for instance, shocks are very delicate to treat in a Lagrangian formalism). Once again, the point is that "there are no shocks" in optimal transport: as discussed in Chapter 8, trajectories do not meet until maybe at final time.

Bibliographical Notes

Otto's seminal paper [290] studied the formal Riemannian structure of the Wasserstein space, and gave applications to the study of the porous medium equation; I shall come back to this topic later. With all the preparations of Part I, the computations performed in this chapter may look rather natural, but they were a little conceptual tour de force at the time of Otto's contribution, and had a strong impact on the research community. This work was partly inspired by the desire to understand in depth a previous contribution by Jordan, Kinderlehrer and Otto [219].

Otto's computations were concerned with the case $U(\rho) = \rho^m$ in $\mathbb{R}^n$, and were generalized later. Otto and the author [292, Section 3] considered $U(\rho) = \rho \log \rho$ on a manifold, and computed the Hessian by differentiating twice along geodesics in the Wasserstein space. To my knowledge, this is the first published work where the Ricci curvature appears in relation to optimal transport.

Later Carrillo, McCann and the author [103] considered functionals of the form $E(\mu) = \int W(x-y)\, \mu(dx)\, \mu(dy)$ in $\mathbb{R}^n$. In my book [365, Section 9.1], I gave formulas for the gradients and Hessians of three basic types of functionals in $\mathbb{R}^n$ that I called *internal energy*, *potential energy* and *interaction energy*, and which can be written respectively (with obvious notation) as

$$\int U(\rho(x))\, dx; \qquad \int V\, d\mu; \qquad \frac{1}{2} \int W(x-y)\, d\mu(x)\, d\mu(y).$$

The functional associated with $U(\rho) = \rho \log \rho$ is Boltzmann's H Functional, well-known in statistical mechanics (see e.g. [364]). The functionals U_ν studied in this chapter (on a manifold and with a reference measure $e^{-V} \mathrm{vol}$) appear in my joint work with Lott [247, Appendix E].

The interpretation of p as a pressure associated to the constitutive law U is well-known in thermodynamics, and was explained to me by McCann; the discussion in the present chapter is slightly expanded in [365, Remarks 5.18].

Presentations of the differential calculus in $P_2(M)$ can be found in my book [365, Chapter 8], and also (with relatively minor variations from one source to the other) in the research papers [292, 98, 294].

Apart from these computations of gradients and Hessians, I am not aware of further studies on Riemannian calculations in $P_2(M)$. The following issues are natural (I am not so sure how useful they would be, but at least they are natural):

- Is there a Jacobi equation in $P_2(M)$, describing small variations of geodesic fields?

- Can one define Christoffel symbols, at least formally?

- Can one define a Laplace operator (taking the trace of the Hessian???)

- Can one define a volume element?? a divergence operator??

The point of view that was first advocated by Otto himself, and which I shall adopt in these notes, is that the "Otto calculus" should primarily be considered a heuristic tool, and the conclusions drawn by its use should then be checked by "direct" means. This might lack elegance, but it is much safer from the point of view of mathematical rigor. Some papers where this strategy has been used with success are [290, 292, 337, 294, 247]. In most of these works, rigorous justifications are mostly done in Lagrangian formalism. The work by Otto and Westdickenberg [294] is an interesting exception to this rule; there everything is attacked from an Eulerian perspective (using such tools as regularization of currents on manifolds).

A different strategy consists in developing an infinite-dimensional Riemannian point of view to make sense of Otto's calculus. Such a theory has been partially developed in [98, 102, 15], at least in the case $M = \mathbb{R}^n$.

An interesting example of functional of the form (15.10), that was considered in relation with optimal transport, is

$$I(\mu) = \int \frac{|\nabla \rho|^2}{\rho}\, dx,$$

which is the Fisher information functional; see [15, Example 11.1.10] and references there provided.

The case $p \neq 2$ is less well understood; as noticed in [15, p. 10], $P_p(M)$ can be seen as a kind of Finsler structure, and there are also rules to compute derivatives in that space, at least to first order. The most general known results to this date are in [15].

16

Displacement convexity I

Convexity plays a prominent role in analysis in general. It is most generally used in a vector space $\mathcal{V}$: A function $F : \mathcal{V} \to \mathbb{R} \cup \{+\infty\}$ is said to be convex if

$$\forall x, y \in \mathcal{V} \quad \forall t \in [0,1] \qquad F\big((1-t)\,x + t\,y\big) \leq (1-t)\,F(x) + t\,F(y). \tag{16.1}$$

But convexity is also a *metric* notion: In short, convexity in a metric space means convexity along geodesics.

In a length space, there always exist geodesic paths joining two given endpoints, so this is a natural setting to define convexity.

Definition 16.1 (convexity in a length space). *Let $(\mathcal{X}, d)$ be a complete length space. Then a function $F : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is said to be geodesically convex, or just convex, if for any constant-speed geodesic path $(\gamma_t)_{0 \leq t \leq 1}$ in $\mathcal{X}$,*

$$\forall t \in [0,1] \qquad F(\gamma_t) \leq (1-t)\,F(\gamma_0) + t\,F(\gamma_1). \tag{16.2}$$

It is said to be weakly convex if for any x_0, x_1 in $\mathcal{X}$ there exists at least one constant-speed geodesic path $(\gamma_t)_{0 \leq t \leq 1}$ with $\gamma_0 = x_0$, $\gamma_1 = x_1$, such that inequality (16.2) *holds true.*

It is a natural problem to identify functionals that are convex on the Wasserstein space. In his 1994 PhD thesis, McCann established and used the convexity of certain functionals on $P_2(\mathbb{R}^n)$ to prove the uniqueness of their minimizers. Since then, his results have been generalized; yet all examples which have been treated so far belong to the general class

$$\mathcal{F}(\mu) = \int_{\mathcal{X}^k} I(x_1, \ldots, x_k)\, d\mu(x_1) \,\ldots\, d\mu(x_k) \;+\; \int_{\mathcal{X}} U\left(\frac{d\mu}{d\nu}\right) d\nu,$$

where $I(x_1, \ldots, x_k)$ is a certain "k-particle interaction potential", U is a nice function $\mathbb{R}_+ \to \mathbb{R}$, and ν is a reference measure.

In this and the next chapter I shall consider the convexity problem on a general Riemannian manifold M, in the case $I = 0$, so the functionals under study will be the functionals U_ν defined by

$$U_\nu(\mu) = \int_M U(\rho)\, d\nu, \qquad \mu = \rho\,\nu. \tag{16.3}$$

In this chapter I shall first give some reminders about the notion of convexity and some of its refinements; then I shall make these notions more explicit in the case of the Wasserstein space $P_2(M)$. In the last section I shall use Otto's calculus to guess sufficient conditions under which U_ν satisfies some interesting convexity properties.

Let the reader not be offended if I strongly insist that convexity in the metric space $P_2(M)$ has nothing to do with the convex structure of the space of probability measures. The former concept will be called "convexity along optimal transport" or "displacement convexity".

Reminders on convexity: differential and integral conditions

The material in this section has nothing to do with optimal transport, and it is, for the most part, rather standard.

It is well-known that a function $F : \mathbb{R}^n \to \mathbb{R}$ is convex, in the sense of (16.1), if and only if it satisfies

$$\nabla^2 F \geq 0 \tag{16.4}$$

(nonnegative Hessian) on $\mathbb{R}^n$. The latter inequality should generally be understood in distribution sense, but let me just forget about this subtlety which is not essential here.

Condition (16.4) is a differential condition, in contrast with the "integral" condition (16.1). There is a more general principle relating a lower bound on the Hessian (differential condition) to a convexity-type inequality (integral condition). It can be stated in terms of the **one-dimensional Green function** (of the Laplace operator with Dirichlet boundary conditions). That Green function is the nonnegative kernel $G(s,t)$ such that for all functions $\varphi \in C([0,1];\mathbb{R}) \cap C^2((0,1);\mathbb{R})$,

$$\varphi(t) = (1-t)\,\varphi(0) + t\,\varphi(1) - \int_0^1 \ddot{\varphi}(s)\, G(s,t)\, ds. \tag{16.5}$$

It is easy to give an explicit expression for G:

$$G(s,t) = \begin{cases} s\,(1-t) & \text{if } s \leq t \\ t\,(1-s) & \text{if } s \geq t \end{cases} \tag{16.6}$$

Then formula (16.5) actually extends to arbitrary continuous functions φ on $[0,1]$, provided that φ'' (taken in distribution sense) is bounded below by a real number.

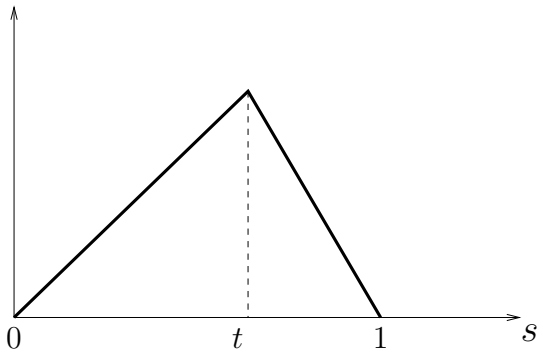


Fig. 16.1. The Green function $G(s,t)$ as a function of s

The next statement provides the equivalence between several differential and integral convexity conditions in a rather general setting.

Proposition 16.2 (lower Hessian bounds). *Let (M, g) be a smooth Riemannian manifold, and let $\Lambda = \Lambda(x, v)$ be a continuous quadratic form on TM; that is, for any x, $\Lambda(x, \cdot)$ is a quadratic form in v, and it depends continuously on x. Assume that $\Lambda \geq Kg$ for some $K \in \mathbb{R}$. Then, for any function $F \in C^2(M)$, the following statements are equivalent:*

(i) $\nabla^2 F \geq \Lambda$

(ii) For any constant-speed, minimizing geodesic path $(\gamma_t)_{0 \leq t \leq 1}$ on M,

$$F(\gamma_t) \leq (1-t)\, F(\gamma_0) + t\, F(\gamma_1) - \int_0^1 \Lambda(\gamma_s, \dot\gamma_s)\, G(s,t)\, ds;$$

(iii) For any constant-speed, minimizing geodesic path $(\gamma_t)_{0 \leq t \leq 1}$ on M,

$$F(\gamma_1) \geq F(\gamma_0) + \big\langle \nabla F(\gamma_0), \dot\gamma_0 \big\rangle + \int_0^1 \Lambda(\gamma_t, \dot\gamma_t)\, (1-t)\, dt.$$

(iv) For any constant-speed, minimizing geodesic path $(\gamma_t)_{0 \leq t \leq 1}$ on M,

$$\big\langle \nabla F(\gamma_1), \dot\gamma_1 \big\rangle - \big\langle \nabla F(\gamma_0), \dot\gamma_0 \big\rangle \geq \int_0^1 \Lambda(\gamma_t, \dot\gamma_t)\, dt.$$

The equivalence is still preserved if conditions (ii), (iii) and (iv) are respectively replaced by the a priori weaker conditions

(ii') For any constant-speed, minimizing geodesic path $(\gamma_t)_{0 \leq t \leq 1}$ on M,

$$F(\gamma_t) \leq (1-t)\, F(\gamma_0) + t\, F(\gamma_1) - \lambda[\gamma]\, \frac{t(1-t)}{2}\, d(\gamma_0, \gamma_1)^2,$$

(iii') For any constant-speed, minimizing geodesic path $(\gamma_t)_{0 \leq t \leq 1}$ on M,

$$F(\gamma_1) \geq F(\gamma_0) + \big\langle \nabla F(\gamma_0), \dot\gamma_0 \big\rangle + \lambda[\gamma]\, \frac{d(\gamma_0, \gamma_1)^2}{2}.$$

(iv') For any constant-speed, minimizing geodesic path $(\gamma_t)_{0 \leq t \leq 1}$ on M,

$$\big\langle \nabla F(\gamma_1), \dot\gamma_1 \big\rangle - \big\langle \nabla F(\gamma_0), \dot\gamma_0 \big\rangle \geq \lambda[\gamma]\, d(\gamma_0, \gamma_1)^2,$$

where

$$\lambda[\gamma] := \inf_{0 \leq t \leq 1} \frac{\Lambda(\gamma_t, \dot\gamma_t)}{|\dot\gamma_t|^2}.$$

Remark 16.3. In the particular case when Λ is equal to λg for some constant $\lambda \in \mathbb{R}$, property (ii) reduces to property (ii') with $\lambda[\gamma] = \lambda$. Indeed, since γ has constant speed,

$$\begin{aligned} F(\gamma_t) &\leq (1-t) F(\gamma_0) + t F(\gamma_1) - \lambda \int_0^1 g(\gamma_s, \dot\gamma_s)\, G(s,t)\, ds \\ &= (1-t) F(\gamma_0) + t F(\gamma_1) - \lambda\, d(\gamma_0, \gamma_1)^2 \int_0^1 G(s,t)\, ds. \end{aligned}$$

By plugging the function $\varphi(t) = t^2$ in (16.5) one sees that $\int_0^1 G(s,t)\, ds = t(1-t)/2$. So (ii) indeed reduces to

$$F(\gamma_t) \leq (1-t) F(\gamma_0) + t F(\gamma_1) - \frac{\lambda\, t(1-t)}{2}\, d(\gamma_0, \gamma_1)^2. \tag{16.7}$$

Definition 16.4 (Λ-convexity). *Let M be a smooth Riemannian manifold, and $\Lambda = \Lambda(x, v)$ a continuous quadratic form on M, $\Lambda \geq Kg$ for some $K \in \mathbb{R}$. Then $F : M \to \mathbb{R} \cup \{+\infty\}$ is said to be Λ-convex if Property (ii) in Proposition 16.2 is satisfied. In the case when $\Lambda = \lambda g$, $\lambda \in \mathbb{R}$, F will be said to be λ-convex; this means that inequality* (16.7) *is satisfied. In particular, 0-convexity is just plain convexity.*

Proof of Proposition 16.2. Assume that (i) holds true. Consider x_0 and x_1 in M, and introduce a constant-speed minimizing geodesic γ joining $\gamma_0 = x_0$ to $\gamma_1 = x_1$. Then

$$\frac{d^2}{dt^2} F(\gamma_t) = \left\langle \nabla^2 F(\gamma_t) \cdot \dot\gamma_t,\, \dot\gamma_t \right\rangle \geq \Lambda(\gamma_t, \dot\gamma_t).$$

Then Property (ii) follows from identity (16.5) with $\varphi(t) := F(\gamma_t)$.

As for Property (iii), it can be established either by dividing the inequality in (ii) by $t > 0$, and then letting $t \to 0$, or directly from (i) by using the Taylor formula at order 2 with $\varphi(t) = F(\gamma_t)$ again. Indeed, $\dot\varphi(0) = \langle \nabla F(\gamma_0),\, \dot\gamma_0 \rangle$, while $\ddot\varphi(t) \geq \Lambda(\gamma_t, \dot\gamma_t)$.

To go from (iii) to (iv), replace the geodesic γ_t by the geodesic γ_{1-t}, to get

$$F(\gamma_0) \geq F(\gamma_1) - \left\langle \nabla F(\gamma_1), \dot\gamma_1 \right\rangle + \int_0^1 \Lambda(\gamma_{1-t}, \dot\gamma_{1-t})\,(1-t)\,dt.$$

After changing variables in the last integral, this is

$$F(\gamma_0) \geq F(\gamma_1) - \left\langle \nabla F(\gamma_1), \dot\gamma_1 \right\rangle + \int_0^1 \Lambda(\gamma_t, \dot\gamma_t)\, t\, dt,$$

and by adding up (iii), one gets Property (iv).

So far we have seen that (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii) $\Rightarrow$ (iv). To complete the proof of equivalence it is sufficient to check that (iv') implies (i).

So assume (iv'). From the identity

$$\left\langle \nabla F(\gamma_1),\, \dot\gamma_1 \right\rangle - \left\langle \nabla F(\gamma_0),\, \dot\gamma_0 \right\rangle = \int_0^1 \nabla^2 F(\gamma_t)(\dot\gamma_t)\, dt,$$

and (iv'), one deduces that, for all geodesic paths γ,

$$\lambda[\gamma]\, d(\gamma_0, \gamma_1)^2 \leq \int_0^1 \nabla^2 F(\gamma_t)(\dot\gamma_t)\, dt. \tag{16.8}$$

Choose (x_0, v_0) in TM, with $v_0 \neq 0$, and $\gamma(t) = \exp_{x_0}(\varepsilon t v_0)$, where $\varepsilon > 0$; of course γ depends implicitly on ε, and $d(\gamma_0, \gamma_1) = \varepsilon |v_0|$. Moreover, as $\varepsilon \to 0$, $(\gamma_t, \dot\gamma_t) \simeq (x_0, \varepsilon v_0)$ in TM, so

$$\lambda[\gamma] = \inf_{0 \leq t \leq 1} \frac{\Lambda(\gamma_t, \dot\gamma_t)}{|\dot\gamma_t|^2} = \inf_{0 \leq t \leq 1} \frac{\Lambda(\gamma_t, \dot\gamma_t/\varepsilon)}{|\dot\gamma_t/\varepsilon|^2} \xrightarrow[\varepsilon \to 0]{} \frac{\Lambda(x_0, v_0)}{|v_0|^2}.$$

So the left-hand side of (16.8) converges to $\Lambda(x_0, v_0)$. On the other hand, since $\nabla^2 F$ is continuous, the right-hand side obviously converges to $\nabla^2 F(x_0)(v_0)$. Then Property (i) follows. □

Displacement convexity

Now I shall discuss convexity in the setting of optimal transport, replacing the manifold M of the previous section by the length space $P_2(M)$. For the moment I shall only consider

measures that are absolutely continuous with respect to the volume on M, and denote by $P_2^{\rm ac}(M)$ the space of such measures. It makes sense to study convexity in $P_2^{\rm ac}(M)$ because this is a *geodesically convex* subset of $P_2(M)$: By Theorem 8.7, a displacement interpolation between any two absolutely continuous measures is itself absolutely continuous. (Singular measures will be considered later, together with singular metric spaces, in Part III of these notes.)

So let μ_0 and μ_1 be two probability measures on M, absolutely continuous with respect to the volume element, and let $(\mu_t)_{0\le t\le 1}$ be the displacement interpolation between μ_0 and μ_1. Recall from Chapter 13 that this displacement interpolation is uniquely defined, and characterized by the formulas $\mu_t = (T_t)_\# \mu_0$, where

$$T_t(x) = \exp_x(t\widetilde{\nabla}\psi(x)), \tag{16.9}$$

and ψ is $d^2/2$-convex. (Forget about the $\widetilde{\ }$ symbol if you don't like it.) Moreover, T_t is injective for $t < 1$; so for all $t < 1$ it makes sense to define the velocity field $v(t,x)$ on $T_t(M)$ by

$$v\big(t, T_t(x)\big) = \frac{d}{dt} T_t(x),$$

and one also has

$$v\big(t, T_t(x)\big) = \widetilde{\nabla}\psi_t(T_t(x)),$$

where ψ_t is a solution at time t of the quadratic Hamilton–Jacobi equation with initial datum $\psi_0 = \psi$.

The next definition adapts the general definitions of convexity, λ-convexity, Λ-convexity. Here λ is a real number that might nonnegative or nonpositive, while $\Lambda = \Lambda(\mu, v)$ defines for each probability measure μ a quadratic form on vector fields $v : M \to TM$; it is assumed that $\Lambda \ge K \int |v|^2 \, d\mu$, for some $K \in \mathbb{R}$.

Definition 16.5 (Displacement convexity). *With the above notation, a functional $F : P_2^{\rm ac}(M) \to \mathbb{R} \cup \{+\infty\}$ is said to be*

- **displacement convex** *if, whenever $(\mu_t)_{0\le t\le 1}$ is a geodesic in $P_2^{\rm ac}(M)$,*

$$\forall t \in [0,1] \qquad F(\mu_t) \le (1-t)\, F(\mu_0) + t\, F(\mu_1);$$

- **λ-displacement convex**, *if, whenever $(\mu_t)_{0\le t\le 1}$ is a geodesic in $P_2^{\rm ac}(M)$,*

$$\forall t \in [0,1] \qquad F(\mu_t) \le (1-t)\, F(\mu_0) + t\, F(\mu_1) - \frac{\lambda\, t(1-t)}{2}\, W_2(\mu_0, \mu_1)^2.$$

- **Λ-displacement convex**, *if, whenever $(\mu_t)_{0\le t\le 1}$ is a geodesic in $P_2^{\rm ac}(M)$, and $(\psi_t)_{0\le t\le 1}$ is an associated solution of the Hamilton–Jacobi equation,*

$$\forall t \in [0,1] \qquad F(\mu_t) \le (1-t)\, F(\mu_0) + t\, F(\mu_1) - \int_0^1 \Lambda(\mu_s, \widetilde{\nabla}\psi_s)\, G(s,t)\, ds,$$

where $G(s,t)$ is the one-dimensional Green function of (16.6).

Of course these definitions are more and more general: Λ-displacement convexity reduces to λ-displacement convexity when $\Lambda(\mu, v) = \lambda \, \|v\|^2_{L^2(\mu)}$; and this in turns reduces to plain displacement convexity when $\lambda = 0$.

Displacement convexity from curvature-dimension bounds

The question is whether such concepts apply to functionals of the form U_ν, as in (16.3). Of course Proposition 16.2 does not apply, because neither $P_2(M)$ nor $P_2^{\rm ac}(M)$ are smooth manifolds. However, if one believes in Otto's formalism, then we can hope that displacement convexity, λ-displacement convexity, Λ-displacement convexity of U_ν would be respectively equivalent to

$$\mathrm{Hess}_\mu U_\nu \geq 0, \qquad \mathrm{Hess}_\mu U_\nu \geq \lambda, \qquad \mathrm{Hess}_\mu U_\nu(\dot\mu) \geq \Lambda(\mu, \dot\mu), \tag{16.10}$$

where $\mathrm{Hess}_\mu U_\nu$ stands for the formal Hessian of U_ν at μ (which was computed in Chapter 15), λ is a shorthand for $\lambda \, \| \cdot \|^2_{L^2(\mu)}$, and $\dot\mu$ is identified with $\nabla\psi$ via the usual continuity equation

$$\dot\mu + \nabla \cdot (\nabla\psi \, \mu) = 0.$$

Now I shall try to identify simple sufficient conditions on the manifold M, the reference measure ν and the energy function U, for (16.10) to hold. This quest is, for the moment, *just formal*; it will be checked later, without any reference to Otto's formalism, that our guess is correct.

To identify conditions for displacement convexity I shall use again the formalism of Chapter 14. Equip the Riemannian manifold M with a reference measure $\nu = e^{-V} \mathrm{vol}$, where V is a smooth function on M, and assume that the resulting space satisfies the curvature-dimension bound $\mathrm{CD}(K, N)$, as in Theorem 14.8, for some $N \in [1, \infty]$ and $K \in \mathbb{R}$. Everywhere in the sequel, ρ will stand for the density of μ with respect to ν.

Consider a continuous function $U : \mathbb{R}_+ \to \mathbb{R}$. I shall assume that U is *convex* and $U(0) = 0$. The latter condition is rather natural from a physical point of view (no matter $\Rightarrow$ no energy). The convexity assumption might seem more artificial, and to justify it I will argue that (i) the convexity of U is necessary for U_ν to be lower semi-continuous with respect to the weak topology induced by the metric W_2; (ii) if one imposes the nonnegativity of the pressure $p(r) = rU'(r) - U(r)$, which is natural from the physical point of view, then conditions for displacement convexity will be in the end quite more stringent than just convexity of U; (iii) the convexity of U automatically implies the nonnegativity of the pressure, since $p(r) = rU'(r) - U(r) = rU'(r) - U(r) + U(0) \geq 0$. For simplicity I shall also impose that U is twice continuously differentiable *everywhere* in $(0, +\infty)$. Finally, I shall assume that ψ in (16.9) is C^2, and I shall avoid the discussion about the domain of definition of U_ν by just considering compactly supported probability measures.

Then, from (15.11) and (14.50),

$$\mathrm{Hess}_\mu U_\nu(\dot\mu) = \int_M \Gamma_2(\psi) \, p(\rho) \, d\nu + \int_M (L\psi)^2 \, p_2(\rho) \, d\nu \tag{16.11}$$

$$\geq \int_M \mathrm{Ric}_{N,\nu}(\nabla\psi) \, p(\rho) \, d\nu + \int_M (L\psi)^2 \, \left[p_2 + \frac{p}{N}\right](\rho) \, d\nu \tag{16.12}$$

$$\geq K \int_M |\nabla\psi|^2 \, p(\rho) \, d\nu + \int_M (L\psi)^2 \, \left[p_2 + \frac{p}{N}\right](\rho) \, d\nu. \tag{16.13}$$

To get a bound on this expression, it is natural to assume that

$$p_2 + \frac{p}{N} \geq 0. \tag{16.14}$$

A typical example is $U = U_N$ defined by

$$U_N(\rho) = \begin{cases} -N\big(\rho^{1-\frac{1}{N}} - \rho\big) & (1 < N < \infty) \\ \\ \rho \log \rho & (N = \infty). \end{cases} \tag{16.15}$$

These functions will come back again and again in the sequel, and the associated functionals will be denoted by $H_{N,\nu}$. The set of all functions U for which (16.14) is satisfied will be called the *displacement convex class of dimension* N and denoted by $\mathcal{DC}_N$.

If inequality (16.14) holds true, then

$$\mathrm{Hess}_\mu\, U_\nu \geq K \Lambda_U,$$

where

$$\Lambda_U(\mu, \dot\mu) = \int_M |\nabla \psi|^2\, p(\rho)\, d\nu. \tag{16.16}$$

So the conclusion is as follows:

Guess 16.6. *Let M be a Riemannian manifold satisfying a curvature-dimension bound* $\mathrm{CD}(K, N)$*, and let U satisfy* (16.14)*; then U_ν is $K\Lambda_U$-displacement convex.*

Note that all the previous discussion makes sense for $N = \infty$.

Actually, there should be an equivalence between the two statements in Guess 16.6. To see this, assume that U_ν is $K\Lambda_U$-displacement convex; pick up an arbitrary point $x_0 \in M$, a tangent vector $v_0 \in T_{x_0} M$, consider the particular function $U = U_N$, a probability measure μ which is very much concentrated close to x_0, and a function ψ such that $\nabla\psi(x_0) = v_0$ and $\Gamma_2(\psi) + (L\psi)^2/N = \mathrm{Ric}_{N,\nu}(v_0)$ (as in the proof of Theorem 14.8). Then, on one hand,

$$K\Lambda_U(\mu, \dot\mu) = K \int |\nabla\psi|^2\, \rho^{1-\frac{1}{N}}\, d\nu \simeq K|v_0|^2 \int \rho^{1-\frac{1}{N}}\, d\nu; \tag{16.17}$$

on the other hand, by the choice of U,

$$\mathrm{Hess}_\mu\, U_\nu(\dot\mu) = \int \left[\Gamma_2(\psi) + \frac{(L\psi)^2}{N}\right] \rho^{1-\frac{1}{N}}\, d\nu,$$

but then since μ is concentrated around x_0, this is well approximated by

$$\left[\Gamma_2(\psi) + \frac{(L\psi)^2}{N}\right](x_0) \int \rho^{1-\frac{1}{N}}\, d\nu = \mathrm{Ric}_{N,\nu}(v_0) \int \rho^{1-\frac{1}{N}}\, d\nu.$$

Comparing that expression with (16.17) shows that $\mathrm{Ric}_{N,\nu}(v_0) \geq K|v_0|^2$. Since x_0 and v_0 were arbitrary, the conclusion is that $\mathrm{Ric}_{N,\nu} \geq K$. Note that this reasoning only used the functional $H_{N,\nu} = (U_N)_\nu$, and probability measures μ that are very concentrated around to a given point.

This heuristic discussion is summarized in the following

Guess 16.7. *If, for each $x_0 \in M$, $H_{N,\nu}$ is $K\Lambda_U$-displacement convex when applied to probability measures that are supported in a small neighborhood of x_0, then M satisfies the* $\mathrm{CD}(K, N)$ *curvature-dimension bound.*

Example 16.8. Condition CD$(0, \infty)$ with $\nu = \text{vol}$ just means Ric ≥ 0, and the statement $U \in \mathcal{DC}_\infty$ just means that the iterated pressure p_2 is nonnegative. The typical example is when $U(\rho) = \rho \log \rho$, and then the corresponding functional is

$$H(\mu) = \int \rho \log \rho \, d\text{vol}\,, \qquad \mu = \rho \, \text{vol}\,.$$

Then the above considerations suggest that the following statements are equivalent:

(i) Ric ≥ 0;

(ii) If the nonlinearity U is such that the nonnegative iterated pressure p_2 is nonnegative, then the functional U_{vol} is displacement convex;

(iii) H is displacement convex;

(iii') For all $x_0 \in M$, the functional H is displacement convex when applied to probability measures that are supported in a small neighborhood of x_0.

Example 16.9. The above considerations also suggest that the inequality Ric $\geq K\, g$ is equivalent to the K-displacement convexity of H, whatever the value of $K \in \mathbb{R}$.

These guesses will be proven and generalized in the next chapter.

A fluid mechanics feeling for Ricci curvature

Ricci curvature is familiar to physicists because it plays a crucial role in Einstein's theory of general relativity. But what we have been discovering in this chapter is that Ricci curvature can also be given a physical interpretation in terms of *classical fluid mechanics*. To give the reader a better feeling of this new point of view, let us imagine how two physicists, the first one used to relativity and light propagation, the second one used to fluid mechanics, would answer the following question: *Describe in an informal way an experiment that can determine whether we live in a nonnegatively Ricci-curved space.*

The Light source test: Take a small light source, and try to determine its volume by looking at it from a distant position. If you systematically overestimate the volume of the light source, then you live in a nonnegatively curved space (recall Figure 14.4).

The Lazy Gas experiment: Take a perfect gas in which particles do not interact, and ask him to move from a certain prescribed density field at time $t = 0$, to another prescribed density field at time $t = 1$. Since the gas is lazy, he will find a way to do so by spending a minimal amount of work (least action path). Measure the entropy of the gas at each time, and check that it always lie *above* the line joining the final and initial entropies. If such is the case, then we know that we live in a nonnegatively curved space.

Bibliographical Notes

Convexity has been extensively studied in the Euclidean space [312] and in Banach spaces [151, 79]. I am not aware of textbooks where the study of convexity in more general length spaces is developed, although this notion is now of rather frequent use (in the context of optimal transport, see e.g. [15, p. 50]).

The concept and terminology of displacement convexity were introduced by McCann in the mid-nineties [267]. He identified Condition (16.14) as the basic criterion for convexity

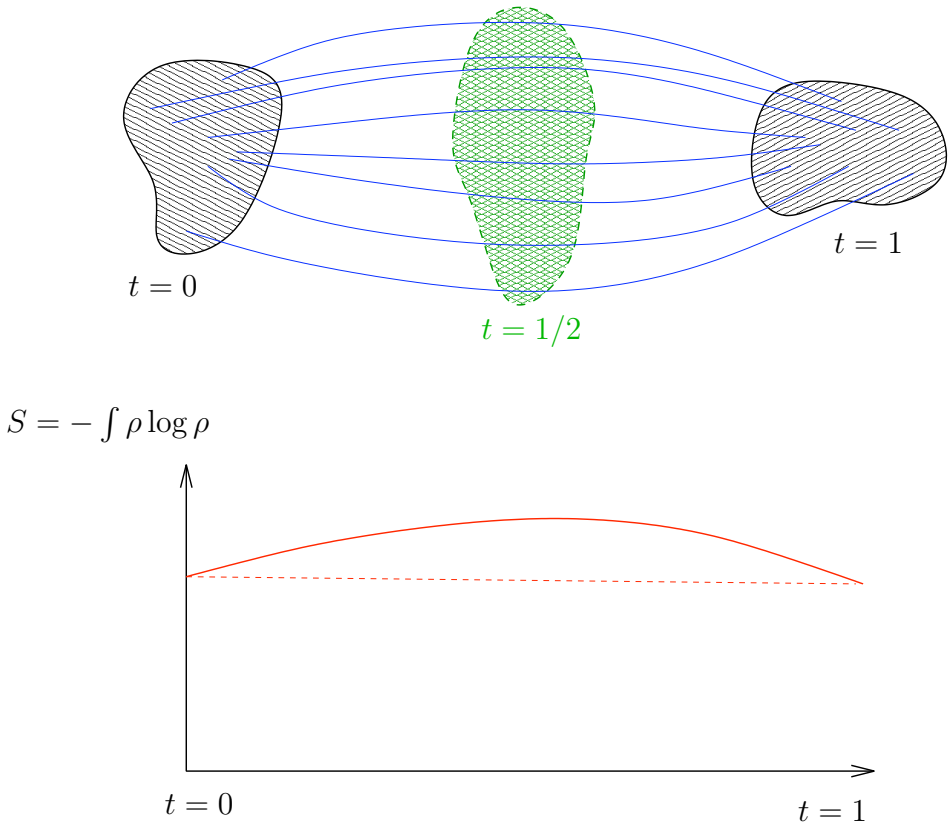


Fig. 16.2. The lazy gas experiment: To go from state 0 to state 1, the lazy gas uses a path of least action. In a nonnegatively curved world, the trajectories of its particles first diverge, then converge, so that at intermediate times he can afford to have a lower density (higher entropy).

in $P_2(\mathbb{R}^n)$, and also discussed other formulations of this condition, that will be studied in the next chapter. Inequality (16.14) was later rediscovered by several authors, in various contexts.

The application of Otto calculus to the study of displacement convexity goes back to [290] and [292]. In the latter reference it was conjectured that nonnegative Ricci curvature would imply displacement convexity of H.

Ricci curvature appears explicitly in Einstein's equations, and will be encountered in any mildly advanced book on general relativity. Fluid mechanics analogies for curvature appear explicitly in the work by Cordero-Erausquin, McCann and Schmuckenschläger [118].

17

Displacement convexity II

In the previous chapter, a conjecture was formulated about the links between displacement convexity and curvature-dimension bounds; the plausibility of this conjecture was justified by some formal computations based on Otto's calculus. Now in the present chapter I shall provide a rigorous justification of this conjecture. In contrast with the previous chapter that was based on an Eulerian point of view, in the present chapter I shall now use a Lagrangian approach. Not only is the Lagrangian formalism easier to justify, but it will also lead to new curvature-dimension criteria ("distorted displacement convexity").

Displacement convexity classes

What I shall call displacement convex class of order N is a family of convex nonlinearities satisfying a certain characteristic differential inequality of second order (recall (16.14)).

Definition 17.1 (displacement convex classes). *Let $N \in [1, \infty]$ be given. Then the class $\mathcal{DC}_N$ is defined as the set of continuous convex functions $U : \mathbb{R}_+ \to \mathbb{R}$, twice continuously differentiable on $(0, +\infty)$, such that $U(0) = 0$, and, with the notation*

$$p(r) = rU'(r) - U(r), \qquad p_2(r) = rp'(r) - p(r),$$

U satisfies any one of the following equivalent differential conditions:

(i) $p_2 + \dfrac{p}{N} \geq 0$;

(ii) $\dfrac{p(r)}{r^{1-1/N}}$ *is a nondecreasing function of* r;

(iii) $u(\delta) := \begin{cases} \delta^N U(\delta^{-N}) & (\delta \in \mathbb{R}_+) \quad \text{if } N < \infty \\ e^{\delta} U(e^{-\delta}) & (\delta \in \mathbb{R}) \quad \text{if } N = \infty \end{cases}$ *is a convex function of* δ.

Remark 17.2. Since U is convex and $U(0) = 0$, the function u appearing in (iii) is automatically nonincreasing.

Remark 17.3. It is clear (from condition (i) for instance) that $\mathcal{DC}_{N'} \subset \mathcal{DC}_N$ for $N' \geq N$. So the smallest class of all is $\mathcal{DC}_\infty$, while $\mathcal{DC}_1$ is the largest (actually, conditions (i)-(iii) are void for $N = 1$).

Remark 17.4. If U belongs to $\mathcal{DC}_N$, then for any $a \geq 0$, $b > 0$, $c \in \mathbb{R}$, the function $r \longmapsto a\, U(br) + cr$ also belongs to $\mathcal{DC}_N$.

Remark 17.5. The requirement for U to be twice differentiable on $(0, +\infty)$ could be removed from most (but probably not all) subsequent results involving displacement convexity classes. Still, this regularity assumption will simplify the proofs, without significantly restricting the generality of applications.

Examples 17.6. (i) For any $\alpha \geq 1$, the function $U(r) = r^\alpha$ belongs to all classes $\mathcal{DC}_N$;

(ii) If $\alpha < 1$, then the function $U(r) = -r^\alpha$ belongs to $\mathcal{DC}_N$ if and only if $N \leq (1-\alpha)^{-1}$ (that is, $\alpha \geq 1 - 1/N$). The function $-r^{1-1/N}$ is in some sense the minimal representative of $\mathcal{DC}_N$.

(iii) The function $U_\infty(r) = r \log r$ belongs to $\mathcal{DC}_\infty$. It can be seen as the limit of the functions $U_N(r) = -N(r^{1-1/N} - r)$, which are the same (up to multiplication and addition of a linear function) than the functions appearing in (ii) above.

Proof of the equivalence in Definition 17.1. Assume first $N < \infty$, and write $r(\delta) = \delta^{-N}$. By computation, $u'(\delta) = -Np(r)/r^{1-1/N}$. So u is convex if and only if $p(r)/r^{1-1/N}$ is a nonincreasing function of δ, i.e. a nondecreasing function of r. Thus (ii) and (iii) are equivalent.

Next, by computation again,

$$u''(\delta) = N^2 r^{\frac{2}{N}-1} \left(p_2(r) + \frac{p(r)}{N} \right). \tag{17.1}$$

So u is convex if and only if $p_2 + p/N$ is nonnegative. This shows the equivalence between (i) and (iii).

In the case $N = \infty$, the arguments are similar, with the formulas

$$r(\delta) = e^{-\delta}, \qquad u'(\delta) = -\frac{p(r)}{r}, \qquad u''(\delta) = \frac{p_2(r)}{r}.$$

□

As noticed in Remark 17.4, linear functions $r \longmapsto c\,r$ lie in all classes $\mathcal{DC}_N$. It is interesting to enquire about the possible superlinear growth of functions in $\mathcal{DC}_N$ at infinity. The following proposition partly answers this question: It shows that nonlinear functions in $\mathcal{DC}_\infty$ grow at least like $r \log r$ at infinity, while for finite N, superlinear functions in $\mathcal{DC}_N$ may grow as slowly as desired.

Proposition 17.7 (asymptotic behavior of functions in $\mathcal{DC}_N$).
(i) If $U \in \mathcal{DC}_\infty$, then either U is linear, or there exists constants $a, b > 0$ such that

$$U(r) \geq a\, r \log r - b\, r.$$

(ii) Let $N \in [1, \infty)$, and let $\Psi \in C(\mathbb{R}_+; \mathbb{R}_+)$ be such that $\Psi(r)/r \longrightarrow \infty$ as $r \to \infty$; then there exists $U \in \mathcal{DC}_N$ such that $0 \leq U \leq \Psi$, and $U(r)/r \longrightarrow \infty$ as $r \to \infty$.

(iii) Let $N \in [1, \infty]$ and let $U \in \mathcal{DC}_N$; then U is the pointwise nondecreasing limit of a sequence of functions $(U_\ell)_{\ell \in \mathbb{N}}$ in $\mathcal{DC}_N$, such that $(U_\ell)'(\infty) < \infty$ and $(U_\ell)'(\infty) \to U'(\infty)$.

(iv) Let $N \in [1, \infty]$ and let $U \in \mathcal{DC}_N$; then U is the pointwise nonincreasing limit of a sequence of functions $(U_\ell)_{\ell \in \mathbb{N}}$ in $\mathcal{DC}_N$, such that $(U_\ell)'(0) > -\infty$, and $(U_\ell)'(0) \to U'(0)$.

Proof of Proposition 17.7. In case (i), the function U can be reconstructed from u by the formula

$$U(r) = r\, u(\log(1/r)), \tag{17.2}$$

As u is convex and nonincreasing, either u is constant (in which case U is linear), or there are constants $a, b > 0$ such that $u(\delta) \geq -a\delta - b$, and then $U(r) \geq -ar \log(1/r) - br = ar \log r - br$.

Now consider property (ii). Without loss of generality, one may assume that Ψ is identically 0 on $[0,1]$ (otherwise, replace Ψ by $\chi\Psi$, where $0 \leq \chi \leq 1$ and χ is identically 0 on $[0,1]$, identically 1 on $[2, +\infty)$). Define a function $u : (0,\infty) \to \mathbb{R}$ by

$$u(\delta) = \delta^N \Psi(\delta^{-N}). \tag{17.3}$$

Then $u \equiv 0$ on $[1,\infty)$, and $\lim_{\delta\to 0^+} u(\delta) = \infty$. The idea is to reconstruct U from u, but the latter might not be convex.

So let $\widetilde{u}$ be the lower convex hull of u on $(0,\infty)$, i.e. the supremum of all linear functions bounded above by u. Then $\widetilde{u} \equiv 0$ on $[1,\infty)$ and $\widetilde{u}$ is nonincreasing. Necessarily,

$$\lim_{\delta\to 0^+} \widetilde{u}(\delta) = +\infty. \tag{17.4}$$

Indeed, suppose on the contrary that $\lim_{\delta\to 0^+} \widetilde{u}(\delta) = M < +\infty$. Let $a \in \mathbb{R}$ be defined by $a := \sup_{\delta\geq 0} \frac{M+1-u(\delta)}{\delta}$ (this function is nonpositive when δ is small enough, so the supremum is finite). Then $u(\delta) \geq M + 1 - a\delta$, so $\lim_{\delta\to 0^+} \widetilde{u}(\delta) \geq M + 1$, which is a contradiction. So (17.4) does hold true.

Now set

$$U(r) := r\, \widetilde{u}(r^{-1/N}). \tag{17.5}$$

Clearly U is continuous and nonnegative, with $U \equiv 0$ on $[0,1]$. As $\widetilde{u}$ is convex and nonincreasing, it follows that U is convex. Hence $U \in \mathcal{DC}_N$. On the other hand, since $\widetilde{u} \leq u$ and $\Psi(r) = r\, u(r^{-1/N})$, it is clear that $U \leq \Psi$; and still (17.4) implies that $U(r)/r$ goes to $+\infty$ as $r \to \infty$.

Next consider statement (iii). To fix the ideas I shall consider the case $N < \infty$, but the case $N = \infty$ works just the same. Let $U \in \mathcal{DC}_N$, and let $u(\delta) = \delta^N U(\delta^{-N})$. We know that u is a nonincreasing, twice differentiable convex function. If u is nonpositive close to 0, this means that U is nonpositive at infinity, and there is nothing to prove; so let us assume that u is positive close to 0. Let then

$$u_\ell(\delta) = \begin{cases} u(\delta) & \text{if } \delta \geq 1/\ell, \\ u(1/\ell) + (\delta - 1/\ell)\, u'(1/\ell) & \text{if } \delta < 1/\ell. \end{cases}$$

It is clear that u_ℓ is still a nonincreasing convex function of δ, bounded above by u, and converges monotonically to u as $\ell \to \infty$. It may fail to be twice differentiable at $\delta = 1/\ell$, but it is not hard to regularize it on the interval $[1/(\ell-1), 1/(\ell+1)]$, in such a way that the regularized function $\widetilde{u}_\ell$ satisfies the same properties as u_ℓ.

Now let

$$U_\ell(r) = r\, \widetilde{u}_\ell(r^{-1/N}).$$

By direct computation,

$$U_\ell''(r) = \frac{r^{-1-\frac{1}{N}}}{N^2} \left(r^{-\frac{1}{N}} \widetilde{u}_\ell''(r^{-\frac{1}{N}}) \,-\, (N-1)\, \widetilde{u}_\ell'(r^{-\frac{1}{N}}) \right).$$

Since $\widetilde{u}_\ell$ is convex nonincreasing, the above expression is nonpositive, and U_ℓ is convex. Then it is easy to check that U_ℓ actually lies in $\mathcal{DC}_N$. On the other hand, as $r \to \infty$, $U_\ell(r)$ is asymptotic to $\widetilde{u}_\ell(0)\, r$, so it has linear growth. This shows that U is the pointwise monotone limit of functions in $\mathcal{DC}_N$ with asymptotic linear growth, as desired.

The proof of (iv) follows the same lines, except that now it is for large values of δ that u should be replaced by a linear function. □

Domain of the functionals U_ν

To any $U \in \mathcal{DC}_N$ corresponds a functional U_ν. However, some conditions might be needed to make sense of $U_\nu(\mu)$. Why is that so? If U is, say, nonnegative, then an integral such as $\int U(\rho)\, d\nu$ always makes sense in $[0, \infty]$, so U_ν is well-defined on the whole of $P_2^{\rm ac}(M)$. But U might be partially negative, and then one should not exclude the possibility that both the negative and the positive parts of $U(\rho)$ have infinite integral. The problem comes from infinity and does not arise on a compact manifold.

Theorem 17.8 below solves this issue: It shows that under some integral condition on ν, the quantity $U_\nu(\mu)$ is well-defined if μ has finite moments of order p large enough. This suggests to study U_ν on the Wasserstein space of order p, rather than on the Wasserstein space of order 2. Since this result only uses the metric structure, I shall state it in the context of general Polish spaces rather than Riemannian manifolds.

Theorem 17.8. *Let $(\mathcal{X}, d)$ be a Polish space and let ν be a Borel measure on $\mathcal{X}$. Let $N \in [1, \infty]$. Assume that there exists $x_0 \in \mathcal{X}$ and $p \in [2, +\infty)$ such that*

$$\begin{cases} \displaystyle\int_{\mathcal{X}} \frac{d\nu(x)}{[1 + d(x_0, x)]^{p(N-1)}} < +\infty & \text{if } N < \infty, \\ \\ \exists c > 0 \qquad \displaystyle\int_M e^{-c\, d(x_0,x)^p}\, d\nu(x) < +\infty & \text{if } N = \infty. \end{cases} \tag{17.6}$$

Then, for all $U \in \mathcal{DC}_N$, the formula

$$U_\nu(\mu) = \int_{\mathcal{X}} U(\rho)\, d\nu, \qquad \mu = \rho\nu$$

unambiguously defines a functional $U_\nu : P_p^{\rm ac}(\mathcal{X}) \to \mathbb{R} \cup \{+\infty\}$, where $P_p^{\rm ac}(\mathcal{X})$ is the set of absolutely continuous probability measures on $\mathcal{X}$ with a finite moment of order p.

Even if no such p exists, U_ν is still well-defined on $P_c^{\rm ac}(\mathcal{X})$, the set of absolutely continuous compactly supported probability measures.

Example 17.9. If ν is the Lebesgue measure on $\mathbb{R}^N$, then U_ν is well-defined on $P_2^{\rm ac}(\mathbb{R}^N)$ for all $U \in \mathcal{DC}_N$, as long as $N \geq 3$. For $N = 2$, Theorem 17.8 allows to define U_ν on $P_p^{\rm ac}(\mathbb{R}^N)$, for any $p > 2$. In the case $N = 1$, U_ν is well-defined on $P_c^{\rm ac}(\mathbb{R}^N)$.

Convention 17.10. In the sequel I shall sometimes write "$p \in [2, +\infty) \cup \{c\}$ satisfying the assumptions of Theorem 17.8". This means that p is either a real number greater or equal than 2, satisfying (17.6) (the metric space $(\mathcal{X}, d)$ and the reference measure ν should be obvious from the context); or the symbol "c", so that $P_p(\mathcal{X})$ stands for the set $P_c(\mathcal{X})$ of compactly supported probability measures.

Remark 17.11. For any positive constant C, the set of probability measures μ in $P_p(\mathcal{X})$ with $\int d(x_0, x)^p\, d\mu(x) \leq C$ is closed in $P_2(\mathcal{X})$; but in general the whole set $P_p(\mathcal{X})$ is not. Similarly, if K is a given compact subset of $\mathcal{X}$, then the set of probability measures with compact support in K is compact in $P_2(\mathcal{X})$; but $P_c(\mathcal{X})$ is not closed in general.

Remark 17.12. If $\mathcal{X}$ is a length space (for instance a Riemannian manifold equipped with its geodesic distance), then $P_p(M)$ is a *geodesically convex* subset of $P_q(M)$, for any $q \in (1, +\infty)$. Indeed, let $(\mu_t)_{0 \leq t \leq 1}$ be a geodesic in $P_q(M)$; according to Corollary 7.20, there is a random geodesic γ such that $\mu_t = \text{law}\,(\gamma_t)$; then the inequalities $\mathbb{E}\, d(x_0, \gamma_0)^p < +\infty$ and $\mathbb{E}\, d(x_0, \gamma_1)^p < +\infty$ together imply $\mathbb{E}\, d(x_0, \gamma_t)^p$, in view of the (crude) inequality

$$0 \le t \le 1 \Longrightarrow\ d(x_0,\gamma_t)^p \le 2^{2p-1}\big[d(x_0,\gamma_0)^p + d(x_0,\gamma_1)^p\big].$$

Combining this with Theorem 8.7, we deduce that $P_p^{\rm ac}(M)$ is geodesically convex in $P_2(M)$, so it is not absurd to study convexity properties of U_ν along geodesics of $P_2(M)$.

Proof of Theorem 17.8. The problem is to show that under the assumptions of the theorem, $U(\rho)$ is bounded below by a ν-integrable function; then $U_\nu(\mu) = \int U(\rho)\,d\nu$ will be well-defined in $\mathbb{R} \cup \{+\infty\}$.

Suppose first that $N < \infty$. By convexity of u, there is a constant $A > 0$ so that $\delta^N\, U(\delta^{-N}) \ge -A\delta - A$, which means

$$U(\rho) \ge -A\big(\rho + \rho^{1-\frac{1}{N}}\big). \tag{17.7}$$

Of course, ρ lies in $L^1(\nu)$; so it is sufficient to show that also $\rho^{1-1/N}$ lies in $L^1(\nu)$. But this is a simple consequence of Hölder's inequality, since

$$\begin{aligned}\int_{\mathcal{X}} \rho(x)^{1-\frac{1}{N}}\,d\nu(x) &= \int_{\mathcal{X}} \big((1+d(x_0,x)^p)\rho(x)\big)^{1-\frac{1}{N}}\,(1+d(x_0,x)^p)^{-1+\frac{1}{N}}\,d\nu(x)\\ &\le \left(\int_{\mathcal{X}} (1+d(x_0,x)^p)\rho(x)\,d\nu(x)\right)^{1-\frac{1}{N}} \left(\int_{\mathcal{X}} (1+d(x_0,x)^p)^{-(N-1)}\,d\nu(x)\right)^{\frac{1}{N}}.\end{aligned}$$

Now suppose that $N = \infty$. By Proposition 17.7(i), there are positive constants A, B such that

$$U(\rho) \ge A\,\rho\log\rho - B\,\rho \tag{17.8}$$

Thus it is sufficient to show that $(\rho\log\rho)_- \in L^1(X,\nu)$. Write

$$\begin{aligned}\int_{\mathcal{X}} \rho(x)\,\log(\rho(x))\,d\nu(x) &= \int_{\mathcal{X}} \rho(x)\,e^{c\,d(x_0,x)^p}\,\log\Big(\rho(x)\,e^{c\,d(x_0,x)^p}\Big)\,e^{-c\,d(x_0,x)^p}\,d\nu(x)\\ &\qquad - c\int_{\mathcal{X}} d(x_0,x)^p\,\rho(x)\,d\nu(x)\end{aligned}$$

$$\begin{aligned}&= \left(\int_{\mathcal{X}} e^{-c\,d(x_0,x)^p}\,d\nu(x)\right)\left(\int_{\mathcal{X}} \rho(x)\,e^{cd(x_0,x)^p}\,\log\Big(\rho(x)\,e^{c\,d(x_0,x)^p}\Big)\,\frac{e^{-c\,d(x_0,x)^p}\,d\nu(x)}{\int_{\mathcal{X}} e^{-c\,d(x_0,x)^p}\,d\nu(x)}\right)\\ &\qquad - c\int_{\mathcal{X}} d(x_0,x)^p\,\rho(x)\,d\nu(x).\end{aligned}$$

Thanks to Jensen's inequality, applied with the probability measure $\frac{e^{-c\,d(x_0,\cdot)^p}\,d\nu}{\int_{\mathcal{X}} e^{-c\,d(x_0,\cdot)^p}\,d\nu}$, the latter expression can be bounded below by

$$\begin{aligned}&\left(\int_{\mathcal{X}} e^{-c\,d(x_0,x)^p}\,d\nu(x)\right)\left(\frac{\int_{\mathcal{X}}\rho\,d\nu}{\int_{\mathcal{X}} e^{-c\,d(x_0,x)^p}\,d\nu(x)}\right)\log\left(\frac{\int_{\mathcal{X}}\rho\,d\nu}{\int_{\mathcal{X}} e^{-c\,d(x_0,x)^p}\,d\nu(x)}\right)\\ &\qquad - c\int_{\mathcal{X}} d(x_0,x)^p\,\rho(x)\,d\nu(x).\end{aligned}$$

This concludes the argument. □

In the sequel of this chapter, I shall study properties of the functionals U_ν, when the base space is a Riemannian manifold M, equipped with its geodesic distance.

Displacement convexity from curvature bounds, revisited

Recall the notation U_N introduced in (16.15) (or in Example 17.6 (iii)). For any $N > 1$, the functional $(U_N)_\nu$ will be rather denoted by $H_{N,\nu}$:

$$H_{N,\nu}(\mu) = \int_M U_N(\rho)\, d\nu, \qquad \mu = \rho\, \nu.$$

I shall often write H_ν instead of $H_{\infty,\nu}$; and I may even write just H if the reference measure is the volume measure. This notation is justified by analogy with the well-known **Boltzmann's H functional** $H(\rho) = \int \rho \log \rho \, d\mathrm{vol}$.

For each $U \in \mathcal{DC}_N$ formula (16.16) defines a functional Λ_U which will later play a role in displacement convexity estimates. It will be convenient to compare this quantity with $\Lambda_N := \Lambda_{U_N}$:

$$\Lambda_N(\mu, v) = \int_M |v(x)|^2\, \rho^{1-\frac{1}{N}}(x)\, d\nu(x).$$

It is clear that $\Lambda_U \geq K_{N,U}\, \Lambda_N$, where

$$K_{N,U} = \inf_{r>0} \frac{K p(r)}{r^{1-1/N}} = \begin{cases} K \lim\limits_{r\to\infty} \dfrac{p(r)}{r^{1-1/N}} & \text{if } K > 0 \\ 0 & \text{if } K = 0 \\ K \lim\limits_{r\to 0} \dfrac{p(r)}{r^{1-1/N}} & \text{if } K < 0. \end{cases} \tag{17.9}$$

It will also be useful to introduce a *local* version of displacement convexity. In short, a functional U_ν is said to be locally displacement convex if it is displacement convex in the neighborhood of each point.

Definition 17.13 (local displacement convexity). *Let M be a Riemannian manifold, and let F be defined on a geodesically convex subset of $P_2^{\mathrm{ac}}(M)$, with values in $\mathbb{R} \cup \{+\infty\}$. Then F is said to be locally displacement convex if, for any $x_0 \in M$ there is $r > 0$ such that the convexity inequality*

$$\forall t \in [0,1] \qquad F(\mu_t) \leq (1-t)\, F(\mu_0) + t\, F(\mu_1)$$

holds true as soon as all measures μ_t, $0 \leq t \leq 1$, are supported in the ball $B_r(x_0)$.

The concepts of local Λ-displacement, and local λ-displacement convexity are defined similarly, by localizing Definition 16.5.

Warning 17.14. When one says that a functional F is locally displacement convex, this *does not* mean that F is displacement convex in a small neighborhood of μ, for any μ. The word "local" refers to the topology of the base space M, not the topology of the Wasserstein space.

The next theorem is a rigorous implementation of Guesses 16.6 and 16.7; it relates curvature-dimension bounds, as appearing in Theorem 14.8, to displacement convexity properties. Recall Convention 17.10.

Theorem 17.15 (curvature-dimension bounds read off from displacement convexity). *Let M be a Riemannian manifold, equipped with its geodesic distance d, and a reference measure $\nu = e^{-V} \mathrm{vol}$, where $V \in C^2(M)$. Let $K \in \mathbb{R}$ and $N \in (1, +\infty]$. Let $p \in [2, +\infty) \cup \{c\}$ satisfy the assumptions of Theorem 17.8. Then the following three conditions are equivalent:*

(i) M satisfies the curvature-dimension criterion $\mathrm{CD}(K, N)$;

(ii) Each $U \in \mathcal{DC}_N$ is $\Lambda_{N,U}$-displacement convex on $P_p^{\mathrm{ac}}(\mathcal{X})$, where $\Lambda_{N,U} = K_{N,U} \Lambda_N$;

(iii) U_N is locally Λ_N-displacement convex.

Remark 17.16. The case $N = 1$ is degenerate since U_1 is not defined; in that case the equivalence (i) $\Leftrightarrow$ (ii) remains true if one defines $K_{N,U}$ to be $+\infty$ if $K > 0$, and 0 if $K \leq 0$. I shall address this case from a slightly different point of view in Theorem 17.32 below. (As stated in that theorem, $N = 1$ is possible only if M is one-dimensional and $\nu = \mathrm{vol}$.)

As a particular case of Theorem 17.15, we now have a rigorous justification of the guess formulated in Example 16.8: nonnegative Ricci curvature is equivalent to the (local) displacement convexity of Boltzmann's H functional. This is the intersection of two situations where Theorem 17.15 is easier to formulate: (a) the case $N = \infty$; and (b) the case $K = 0$. These cases are important enough to be stated explicitly as corollaries of Theorem 17.15:

Corollary 17.17 ($\mathrm{CD}(K, \infty)$ and $\mathrm{CD}(0, N)$ bounds via optimal transport). *Let (M, g) be a Riemannian manifold, $K \in \mathbb{R}$ and $N \in (1, \infty]$; then*

(a) M satisfies $\mathrm{Ric} \geq K\, g$ if and only if Boltzmann's functional H is K-displacement convex on $P_c^{\mathrm{ac}}(M)$;

(b) M has nonnegative Ricci curvature and dimension bounded above by N if and only if $H_{N,\mathrm{vol}}$ is displacement convex on $P_c^{\mathrm{ac}}(M)$.

Remark 17.18. All these results can be extended to singular measures, so the restriction to absolutely continuous measures is nonessential. I shall come back to these issues in Part III of these notes.

Core of the proof of Theorem 17.15. Before giving a complete proof, I shall give the main argument for the implication (i) $\Rightarrow$ (ii) in Theorem 17.15, in the simple case $K = 0$.

Let $(\mu_t)_{0 \leq t \leq 1}$ be a Wasserstein geodesic, with μ_t absolutely continuous, and let ρ_t be the density of μ_t with respect to ν. It will follow by change of variables that

$$\int U(\rho_t)\, d\nu = \int U\left(\frac{\rho_0}{\mathcal{J}_t}\right) \mathcal{J}_t\, d\nu,$$

where $\mathcal{J}_t$ is the Jacobian of the optimal transport taking μ_0 to μ_t. The next step consists in rewriting this as a function of the mean distortion. Let $u(\delta) = \delta^N U(\delta^{-N})$, then

$$\int U\left(\frac{\rho_0}{\mathcal{J}_t}\right) \mathcal{J}_t\, d\nu = \int u\left(\frac{\mathcal{J}_t^{\frac{1}{N}}}{\rho_0^{\frac{1}{N}}}\right) \rho_0\, d\nu.$$

The fact that U belongs to $\mathcal{DC}_N$ means precisely that u is convex nonincreasing. The nonnegativity of Ricci curvature means that the expression inside the brackets is a concave function of t. Then the convexity of the whole expression follows from the simple fact that the composition of a convex nonincreasing function with a concave function is itself convex. □

Complete proof of Theorem 17.15.

Let us start with the proof of (i) $\Rightarrow$ (ii). I shall only treat the case $N < \infty$, since the case $N = \infty$ is very similar. In a first step, I shall also assume that μ_0 and μ_1 are compactly supported; this assumption will be relaxed in a second step.

So let μ_0 and μ_1 be two absolutely continuous, compactly supported probability measures and let $(\mu_t)_{0\le t\le 1}$ be the unique displacement interpolation between μ_0 and μ_1. It can be written $(T_t)_\# \mu_0$, where $T_t(x) = \exp_x(t\nabla\psi(x))$, let then $(\psi_t)_{0\le t\le 1}$ solve the Hamilton–Jacobi equation with initial datum $\psi_0 = \psi$. The goal is to show that

$$U_\nu(\mu_t) \le (1-t)\, U_\nu(\mu_0) + t\, U_\nu(\mu_1) - K_{N,U} \int_0^1 \int_M \rho_s(x)^{1-\frac{1}{N}}\, |\nabla\psi_s(x)|^2\, d\nu(x)\, G(s,t)\, ds. \tag{17.10}$$

If either $U_\nu(\mu_0) = \infty$ or $U_\nu(\mu_1) = \infty$, then there is nothing to prove; so let us assume that these quantities are finite.

Let t_0 be a fixed time in $(0,1)$; on $T_{t_0}(M)$, define, for all $t \in [0,1]$,

$$T_{t_0\to t}\big(\exp_x(t_0\nabla\psi(x))\big) = \exp_x(t\nabla\psi(x)).$$

Then $T_{t_0\to t}$ is the unique optimal transport $\mu_{t_0} \to \mu_t$. Let $\mathcal{J}_{t_0\to t}$ be the associated Jacobian determinant (well-defined μ_{t_0}-almost surely). Recall from Chapter 11 that μ_t is concentrated on $T_{t_0\to t}(M)$ and that its density ρ_t is determined by the equation

$$\rho_{t_0}(x) = \rho_t(T_{t_0\to t}(x))\, \mathcal{J}_{t_0\to t}(x). \tag{17.11}$$

Since $U(0) = 0$, it is possible to apply Theorem 11.3 to $F(x) = U(\rho_t(x))$; or more precisely, to the positive part and the negative part of U separately. So

$$\int_M U(\rho_t(x))\, d\nu(x) = \int_M U\big(\rho_t(T_{t_0\to t}(x))\big)\, \mathcal{J}_{t_0\to t}(x)\, d\nu(x).$$

Then formula (17.11) implies

$$\int_M U(\rho_t)\, d\nu = \int_M U\left(\frac{\rho_{t_0}(x)}{\mathcal{J}_{t_0\to t}(x)}\right)\, \mathcal{J}_{t_0\to t}(x)\, d\nu(x). \tag{17.12}$$

Since the contribution of $\{\rho_{t_0} = 0\}$ does not matter, this can be rewritten

$$\begin{aligned} U_\nu(\mu_t) &= \int_M U\left(\frac{\rho_{t_0}(x)}{\mathcal{J}_{t_0\to t}(x)}\right)\, \frac{\mathcal{J}_{t_0\to t}(x)}{\rho_{t_0}(x)}\, \rho_{t_0}(x)\, d\nu(x) \\ &= \int_M U\Big(\delta_{t_0}(t,x)^{-N}\Big)\, \delta_{t_0}(t,x)^N\, d\mu_{t_0}(x) \\ &= \int_M w(t,x)\, d\mu_{t_0}(x), \end{aligned}$$

where $w(t,x) := U(\delta_{t_0}(t,x)^{-N})\, \delta_{t_0}(t,x)^N$, and

$$\delta_{t_0}(t,x) = \rho_t\big(T_{t_0\to t}(x)\big)^{-\frac{1}{N}} = \left(\frac{\mathcal{J}_{t_0\to t}(x)}{\rho_{t_0}(x)}\right)^{\frac{1}{N}}.$$

Up to a factor which does not depend on t, $\delta_{t_0}(\cdot, x)$ coincides with $\mathcal{D}(t)$ in the notation of Chapter 14. So, by Theorem 14.8, for almost all x one has

$$\ddot{\delta}_{t_0}(t,x) \le -\frac{K}{N}\, \delta_{t_0}(t,x)\, \big|\nabla\psi_t(T_{t_0\to t}(x))\big|^2.$$

Set $u(\delta) = \delta^N\, U(\delta^{-N})$, so that $w = u \circ \delta$, where δ is a shorthand for $\delta_{t_0}(\cdot, x)$ and x is fixed. Then, since u is convex and $\dot u = -N p(r)/r^{1-1/N} \le 0$, one has, with $r = \delta^{-N}$,

$$\frac{\partial^2 w}{\partial t^2} = \left(\frac{\partial^2 u}{\partial \delta^2}\right) (\dot\delta(t))^2 + \left(\frac{\partial u}{\partial \delta}\right) \ddot\delta(t) \ge \left(-N \frac{p(r)}{r^{1-\frac{1}{N}}}\right) \left(-\frac{K}{N}\, \delta(t)\, |\nabla \psi_t(T_{t_0 \to t}(x))|^2\right).$$

By combining this with the definition of $K_{N,U}$, one obtains

$$\ddot w(t,x) \ge K_{N,U}\, \delta_{t_0}(t,x) = K_{N,U}\, \rho_t(T_{t_0 \to t}(x))^{-\frac{1}{N}} |\nabla \psi_t(T_{t_0 \to t}(x))|^2. \tag{17.13}$$

Since w is a continuous function of t, this implies (recall Proposition 16.2)

$$w(t,x) - (1-t) w(0,x) - t w(1,x) \le -K_{N,U} \int_0^1 \rho_s(T_{t_0 \to s}(x))^{-\frac{1}{N}} \left|\nabla \psi_s(T_{t_0 \to s}(x))\right|^2 G(s,t)\, ds.$$

Upon integration against μ_{t_0}, this inequality becomes

$$\begin{aligned}
U_\nu(\mu_t) - (1-t)\, U_\nu(\mu_0) - t\, U_\nu(\mu_1) &\le -K_{N,U} \int_M \left(\int_0^1 \rho_s(T_{t_0 \to s}(x))^{-\frac{1}{N}}\, |\nabla \psi_s(T_{t_0 \to s}(x))|^2\, G(s,t)\, ds \right) d\mu_{t_0}(x) \\
&= -K_{N,U} \int_0^1 \int_M \rho_s(T_{t_0 \to s}(x))^{-\frac{1}{N}}\, |\nabla \psi_s(T_{t_0 \to s}(x))|^2\, d\mu_{t_0}(x)\, G(s,t)\, ds \\
&= -K_{N,U} \int_0^1 \int_M \rho_s(y)^{-\frac{1}{N}}\, |\nabla \psi_s(y)|^2\, d\mu_s(y)\, G(s,t)\, ds \\
&= -K_{N,U} \int_0^1 \int_M \rho_s(y)^{1-\frac{1}{N}}\, |\nabla \psi_s(y)|^2\, d\nu(y)\, G(s,t)\, ds.
\end{aligned}$$

This concludes the proof of Property (ii) when μ_0 and μ_1 have compact support. Now the general case will be obtained by a restriction argument. Let $p \in [2, +\infty)$ satisfy the assumptions of Theorem 17.8, and let μ_0, μ_1 be two probability measures in $P_p^{\rm ac}(M)$. Let $(Z_\ell)_{\ell \in \mathbb{N}}$, $(\mu_{t,\ell})_{0 \le t \le 1,\, \ell \in \mathbb{N}}$ $(\psi_{t,\ell})_{0 \le t \le 1,\, \ell \in \mathbb{N}}$ be as in Proposition 13.2. Let $\rho_{t,\ell}$ stand for the density of $\mu_{t,\ell}$. By Remark 17.4, the function $U_\ell : r \to U(Z_\ell r)$ belongs to $\mathcal{DC}_N$; and it is easy to check that $K_{N,U_\ell} = K_{N,U}$. Since the $\mu_{t,\ell}$ are compactly supported, we can apply the previous inequality with μ_t replaced by $\mu_{t,\ell}$ and U replaced by U_ℓ:

$$\begin{aligned}
\int U(Z_\ell\, \rho_{t,\ell})\, d\nu \le (1-t) \int U(Z_\ell\, \rho_{0,\ell})\, d\nu + t \int U(Z_\ell\, \rho_{1,\ell})\, d\nu \\
- K_{N,U} \int_0^1 \int_M \rho_{s,\ell}(y)^{1-\frac{1}{N}} |\nabla \psi_{s,\ell}(y)|^2\, d\nu(y)\, G(s,t)\, ds.
\end{aligned} \tag{17.14}$$

It remains to pass to the limit in (17.14) as $\ell \to \infty$. Recall from Proposition 13.2 that $Z_\ell\, \rho_{t,\ell}$ is a nondecreasing family of functions converging monotonically to ρ_t. Since U_+ is nondecreasing, it follow that

$$U_+(Z_\ell\, \rho_{t,\ell}) \;\uparrow\; U_+(\rho_t).$$

On the other hand, the proof of Theorem 17.8 shows that $U_-(r) \le A(r + r^{1-\frac{1}{N}})$ for some $A = A(N,U)$; so

$$U_-(Z_\ell\, \rho_{t,\ell}) \le A\big(Z_\ell\, \rho_{t,\ell} + Z_\ell^{1-\frac{1}{N}}\, \rho_{t,\ell}^{1-\frac{1}{N}}\big) \le A\left(\rho_t + \rho_t^{1-\frac{1}{N}}\right). \tag{17.15}$$

By the proof of Theorem 17.8 and Remark 17.12, the function on the right-hand side of (17.15) is ν-integrable. As a conclusion,

$$\int U_+(Z_\ell\,\rho_{t,\ell})\,d\nu \xrightarrow[\ell\to\infty]{} \int U_+(\rho_t)\,d\nu \qquad \text{by monotone convergence;}$$

$$\int U_-(Z_\ell\,\rho_{t,\ell})\,d\nu \xrightarrow[\ell\to\infty]{} \int U_-(\rho_t)\,d\nu \qquad \text{by dominated convergence.}$$

So we can pass to the limit in the first three terms appearing in the inequality (17.14). As for the last term, note that $|\nabla\psi_{s,\ell}(y)|^2 = d(y, T_{s\to 1,\ell}(y))^2$, at least $\mu_{s,\ell}(dy)$-almost surely; but then this coincides with $d(y, T_{s\to 1}(y))^2$, according to Proposition 13.2. So the last integral in (17.14) can be rewritten as

$$\frac{1}{Z_\ell^{1-\frac{1}{N}}} \int_0^1 \int_M (Z_\ell\,\rho_{s,\ell}(y))^{1-\frac{1}{N}}\, d(y, T_{s\to 1})^2\, d\nu(y)\, G(s,t)\, ds,$$

and by monotone convergence this goes to

$$\int_0^1 \int_M (\rho_s(y))^{1-\frac{1}{N}}\, d(y, T_{s\to 1})^2\, d\nu(y)\, G(s,t)\, ds$$

as $k \to \infty$. So we have passed to the limit in all terms of (17.14), and the proof of (i) $\Rightarrow$ (ii) is complete.

Since the implication (ii) $\Rightarrow$ (iii) is trivial, to conclude the proof of Theorem 17.15 it only suffices to prove that (iii) implies (i). So let $x_0 \in M$; the goal is to show that $(\mathrm{Ric}_{N,\nu})_{x_0} \geq K\, g_{x_0}$, where g is the Riemannian metric. Let $r > 0$ be such that $H_{N,\nu}$ is Λ_N-displacement convex in $B_r(x_0)$. Let $v_0 \neq 0$ be a tangent vector at x_0. As in the proof of Theorem 14.8, we can construct $\widetilde{\psi} \in C^2(M)$, compactly supported in $B_r(x_0)$, such that $\nabla\widetilde{\psi}(x_0) = v_0$, $\nabla^2\widetilde{\psi}(x_0) = \lambda_0\, I_n$ (where I_n is the identity on $T_{x_0}M$) and

$$\left[\Gamma_2(\widetilde{\psi}) + \frac{(L\widetilde{\psi})^2}{N}\right](x_0) = \mathrm{Ric}_{N,\nu}(v_0).$$

Let then $\psi := \theta\widetilde{\psi}$, where θ is a positive real number. If θ is small enough, then ψ is $d^2/2$-convex by Theorem 13.4, and $|\nabla\psi| \leq r/2$. Let then ρ_0 be a smooth probability density, supported in $B_\eta(x_0)$, with $\eta < r/2$. Define

$$\mu_0 = \rho_0\,\nu; \qquad \mu_t = \exp(t\nabla\psi)_\#\mu_0.$$

Then $(\mu_t)_{0\leq t\leq 1}$ is a geodesic in $P_2(M)$, and it is entirely supported in $B_r(x_0)$, so condition (iii) implies

$$H_{N,\nu}(\mu_t) - (1-t)\,H_{N,\nu}(\mu_0) - t\,H_{N,\nu}(\mu_1) \geq K \int_0^1 \left(\int \rho_s(x)^{1-\frac{1}{N}}\, |\nabla\psi_s(x)|^2\, d\nu(x)\right) ds. \tag{17.16}$$

As in the proof of (i) $\Rightarrow$ (ii), let $\delta(t,x)$ be the Jacobian determinant of the map $\exp(t\nabla\psi)$ at x. (This amounts to choose $t_0 = 0$ in the computations above; now this is not a problem since $\exp(t\nabla\psi)$ is for sure Lipschitz.) Let further $\gamma(t,x) = \exp_x(t\nabla\psi(x))$. Formula (14.39) becomes

$$\begin{aligned} -N\frac{\ddot{\delta}(t,x)}{\delta(t,x)} = \mathrm{Ric}_{N,\nu}(\dot\gamma(t,x)) &+ \left\| U(t,x) - \left(\frac{\mathrm{tr}\; U(t,x)}{n}\right) I_n \right\|_{\mathrm{HS}}^2 \\ &+ \frac{n}{N(N-n)} \left[\left(\frac{N-n}{n}\right) \mathrm{tr}\; U(t,x) + \dot\gamma(t,x)\cdot\nabla V(\gamma(t,x))\right]^2, \end{aligned} \tag{17.17}$$

where $U(t,x)$ solves $U(0,x) = \nabla^2\psi(x)$ and U is the solution of the differential equation $\dot{U} + U^2 + R = 0$, where R is defined by (14.7). By using all this information, we shall derive expansions of (17.17) as $\theta \to 0$, $\widetilde{\psi}$ being fixed. First of all, $x = x_0 + O(\theta)$ (this is formal writing to mean that $d(x,x_0) = O(\theta)$; then, by smoothness of the exponential map, $\dot{\gamma}(t,x) = \theta v_0 + O(\theta^2)$; it follows that $\mathrm{Ric}_{N,\nu}(\dot{\gamma}(t,x)) = \theta^2 \mathrm{Ric}_{N,nu;x_0}(v_0) + O(\theta^3)$. Next, $U(0) = \theta\nabla^2\widetilde{\psi}(x_0) = \lambda_0\theta I_n$; so the differential inequality satisfied by U leads to $\dot{U}(0) = -(\lambda_0\theta)^2 I_n - R(0) = O(\theta^2)$, so $U(t,x) = \lambda_0\theta I_n + O(\theta^2)$. Also $U - (\mathrm{tr}\, U) I_n/n = O(\theta^2)$, $\mathrm{tr}\, U(t) = \lambda_0\theta n + O(\theta^2)$ and $\dot{\gamma}(t,x)\cdot\nabla V(\gamma(t,x)) + ((N-n)/n)\,\mathrm{tr}\, U(t,x) = O(\theta^2)$. Plugging all these expansions in (17.17), we get

$$\frac{\ddot{\delta}(t,x)}{\delta(t,x)} = \frac{1}{N}\Big(-\theta^2\,\mathrm{Ric}_{N,\nu}(v_0) + O(\theta^3)\Big). \tag{17.18}$$

By repeating the proof of (i) $\Rightarrow$ (ii) with $U = U_N$ and using (17.18), one obtains

$$\begin{aligned} H_{N,\nu}(\mu_t) - (1-t)\,H_{N,\nu}(\mu_0) - t\,H_{N,\nu}(\mu_1) \\ \geq -\theta^2\big(\mathrm{Ric}_{N,\nu}(v_0) + O(\theta)\big) \int_0^1 \int_M \rho_s(y)^{1-\frac{1}{N}}\, d\nu(y)\, G(s,t)\, ds. \end{aligned} \tag{17.19}$$

On the other hand, by assumption

$$\begin{aligned} H_{N,\nu}(\mu_t) - (1-t)\,H_{N,\nu}(\mu_0) - t\,H_{N,\nu}(\mu_1) &\leq -K \int_0^1 \int_M \rho_s(y)^{1-\frac{1}{N}}\, |\dot{\gamma}(s,y)|^2\, d\nu(y)\, G(s,t)\, ds \\ &= -K\,\theta^2(|v_0|^2 + O(\theta)) \int_0^1 \int_M \rho_s(y)^{1-\frac{1}{N}}\, |\dot{\gamma}(s,y)|^2\, d\nu(y)\, G(s,t)\, ds. \end{aligned}$$

(The constant K is here because $K_{N,U_N} = K$.) By combining this with (17.19) and cancelling out the factors $\theta^2 \int_0^1 \int \rho_s(y)^{1-\frac{1}{N}}\, d\nu(y)\, G(s,t)\, ds$, one concludes that

$$\mathrm{Ric}_{N,\nu}(v_0) \geq K|v_0|^2 + O(\theta).$$

The conclusion follows upon taking the limit $\theta \to 0$. □

Exercise 17.19 (Necessary condition for displacement convexity). This exercise shows that elements of $\mathcal{DC}_N$ are essentially the only candidates for displacement convexity. Let N be a positive integer, $M = \mathbb{R}^N$, and let ν be the Lebesgue measure in $\mathbb{R}^N$. Let U be a measurable function $\mathbb{R}_+ \to \mathbb{R}$ such that U_ν is lower semi-continuous and convex on the space $P_c^{ac}(\mathbb{R}^N)$ (absolutely continuous, compactly supported probability measures), equipped with the distance W_2. Show that (a) U is convex lower semi-continuous; (b) $\delta \to \delta^N\, U(\delta^{-N})$ is convex. Hint: To prove (b), consider the geodesic curve $(\mu_\delta)_{\delta>0}$, where μ_δ is the uniform probability measure on $B_\delta(0)$.

Exercise 17.20 (a bit tricky). Show that if (M,ν) satisfies $\mathrm{CD}(K,N)$ and $U \in \mathcal{DC}_N$, then U_ν is $K\lambda$-displacement convex, when restricted to the geodesically convex set defined by

$$\rho \leq R,$$

with $\lambda = R^{1/N}$. In short, U_ν is $K\|\rho\|_{L^\infty}^{1/N}$-displacement convex.

Ricci curvature bounds from distorted displacement convexity

In Theorem 17.15, all the influence of the Ricci curvature bounds lies in the additional term $\int_0^1 (\dots)\, G(s,t)\, ds$. As a consequence, as soon as $K \neq 0$ and $N < \infty$, the formulation involves not only μ_t, μ_0 and μ_1, but the whole geodesic path $(\mu_s)_{0 \leq s \leq 1}$. This makes the exploitation of the resulting inequality (in geometric applications, for instance) somewhat delicate, if not impossible.

This difficulty was elegantly solved by Sturm, who suggested a different formulation, expressed only in terms of μ_t, μ_0 and μ_1; but now, the functionals $U_\nu(\mu_0)$ and $U_\nu(\mu_1)$ are replaced by other expressions of μ_0 in which extra distortion coefficients will appear. From the technical point of view, this new formulation relies on the principle that one can "take the direction of motion out", in all reformulations of Ricci curvature bounds that were examined in Chapter 14. Before stating the definition of distorted displacement convexity, I shall spend some time on the distorted U_ν functionals.

Definition 17.21 (Distorted U_ν functional). *Let $(\mathcal{X}, d)$ be a Polish space equipped with a reference measure ν. Let U be a convex function with $U(0) = 0$, and let β be a positive continuous function on $\mathcal{X} \times \mathcal{X}$. The distorted U_ν functional with distortion coefficient β is defined as follows: For any probability measure π on $\mathcal{X} \times \mathcal{X}$, whose marginal μ on the first factor is absolutely continuous,*

$$U^\beta_{\pi,\nu}(\mu) = \int_{\mathcal{X}\times\mathcal{X}} U\left(\frac{\rho(x)}{\beta(x,y)}\right) \beta(x,y)\, \pi(dy|x)\, \nu(dx), \tag{17.20}$$

where $\pi(dy|x)$ is the disintegration of π with respect to its first marginal. In particular, if π is a deterministic transference plan associated with some transport $T : \mathcal{X} \to \mathcal{Y}$, then

$$U^\beta_{\pi,\nu}(\mu) = \int_{\mathcal{X}} U\left(\frac{\rho(x)}{\beta(x,T(x))}\right) \beta(x,T(x))\, \nu(dx), \tag{17.21}$$

Remark 17.22. Most of the time, we shall use Definition 17.20 with $\beta = \beta_t^{(K,N)}$, that is, the *reference distortion coefficients* introduced in Definition 14.19.

The same problems of domain of definition which we encountered for the original U_ν functionals arise for the distorted ones; the next theorem solves this issue.

Theorem 17.23 (Domain of definition of U^β_ν). *Let $(\mathcal{X}, d)$ be a Polish space, equipped with a Borel measure ν; let $K \in \mathbb{R}$ and $N \in [1, +\infty]$. Let further $U \in \mathcal{DC}_N$, let β be a continuous positive function on $\mathcal{X} \times \mathcal{X}$, and π be a probability measure on $\mathcal{X} \times \mathcal{X}$, such that the marginal μ of π is absolutely continuous with density ρ.*

Let $p \in [2, +\infty)$ be such that

$$\begin{cases} \displaystyle\int_{\mathcal{X}} \frac{d\nu(x)}{[1 + d(\star, x)]^{p(N-1)}} < +\infty & (N < \infty), \\ \exists c > 0 \qquad \int_{\mathcal{X}} e^{-c\, d(\star,x)^p}\, d\nu(x) < +\infty & (N = \infty). \end{cases} \tag{17.22}$$

If $\mathcal{X}$ is not compact, further assume that β satisfies the following bounds:

$$\begin{cases} \beta \text{ is bounded} & (N < \infty) \\ \displaystyle\int_{\mathcal{X}\times\mathcal{X}} (\log \beta(x,y))_+\, \pi(dx\, dy) < +\infty & (N = \infty). \end{cases} \tag{17.23}$$

Then the integral $U^\beta_{\pi,\nu}(\mu)$ appearing in Definition 17.21 makes sense in $\mathbb{R} \cup \{+\infty\}$ as soon as $\mu \in P_p^{\rm ac}(\mathcal{X})$.

Even if there is no such p, $U^\beta_{\pi,\nu}(\mu)$ still makes sense if $\mu \in P_c^{\rm ac}(\mathcal{X})$.

Remark 17.24. When $\mathcal{X}$ is a Riemannian manifold M satisfying the $\mathrm{CD}(K,N)$ curvature-dimension bounds, then the distortion coefficients $\beta_t^{(K,N)}$ satisfy the assumptions of Theorem 17.23. Indeed,

- If $K \le 0$, then $\beta_t^{(K,N)}$ is bounded;
- If $K > 0$ and $N < \infty$, then M is compact (by the Bonnet–Myers theorem);
- If $K > 0$ and $N = \infty$, then $\log \beta_t^{(K,N)}(x,y)$ is bounded above by a constant multiple of $d(x,y)^2$, which is $\pi(dx\,dy)$-integrable whenever π is an optimal coupling arising in some displacement interpolation.

Proof of Theorem 30.4. The argument is similar to the proof of Theorem 17.8. In the case $N < \infty$, it suffices to write

$$\beta U(\rho/\beta) \ge -A\beta\Big(\frac{\rho}{\beta} + \Big(\frac{\rho}{\beta}\Big)^{1-\frac{1}{N}}\Big) = -A\rho - A\beta^{\frac{1}{N}}\rho^{1-\frac{1}{N}};$$

then the right-hand side is integrable since $\rho^{1-1/N}$ is integrable (as noted in the proof of Theorem 17.8) and β is bounded.

In the case $N = \infty$, the proof of Theorem 17.8 shows that $\rho(\log\rho)_- \in L^1(\nu)$, and $U(\rho) \ge A\rho\log\rho - B\rho$ for some positive constants A, B; so $U_\nu(\mu)$ is well-defined in $\mathbb{R}\cup\{+\infty\}$. Also

$$\begin{aligned}\beta(x,y) U\left(\frac{\rho(x)}{\beta(x,y)}\right) &\ge A\beta(x,y)\left(\frac{\rho(x)}{\beta(x,y)}\right)\log\left(\frac{\rho(x)}{\beta(x,y)}\right) - B\rho(x)\\ &= A\rho(x)\log\rho(x) - A\rho(x)\log\beta(x,y) - B\rho(x).\end{aligned}$$

We already know that the first and third terms in the right-hand side are integrable against $\pi(dy|x)\,\nu(dx)$; as for the second one, it is also integrable since

$$\begin{aligned}\int \rho(x)\,(\log\beta(x,y))_+\,\pi(dy|x)\,\nu(dx) &\le \int (\log\beta(x,y))_+ \pi(dy|x)\,\mu(dx)\\ &= \int (\log\beta(x,y))_+\,\pi(dx\,dy).\end{aligned}$$

This concludes the proof of Theorem 17.23. □

Now comes the key notion in this section:

Definition 17.25 (Distorted displacement convexity). *Let M be a Riemannian manifold, equipped with a reference measure ν. Let $(\beta_t(x,y))_{0\le t\le 1}$ be a family of nonnegative functions on $M\times M$, and let $U : \mathbb{R}_+ \to \mathbb{R}$ be a continuous convex function defined on a geodesically convex subset of $P_2^{\rm ac}(M)$. The functional U_ν is said to be displacement convex with distortion (β_t) if, for all geodesic path $(\mu_t)_{0\le t\le 1}$ in the domain of U_ν,*

$$\forall t \in [0,1], \qquad U_\nu(\mu_t) \le (1-t)\,U^\beta_{\pi,\nu}(\mu_0) \;+\; t\,U^\beta_{\check\pi,\nu}(\mu_1), \tag{17.24}$$

where π stands for the optimal transference plan between μ_0 and μ_1; and $\check\pi$ is obtained from π by switching the variables, that is $\check\pi = S_\#\pi$, $S(x_0,x_1) = (x_1,x_0)$.

This notion can be localized as in Definition 17.13.

Remark 17.26. The inequality appearing in (17.24) can be rewritten more explicitly as

$$\int U(\rho_t)\,d\nu \le (1-t)\int_{M\times M} U\left(\frac{\rho_0(x_0)}{\beta_{1-t}(x_0,x_1)}\right)\beta_{1-t}(x_0,x_1)\,\pi(dx_0|x_1)\,\nu(dx_0) + t\int_{M\times M} U\left(\frac{\rho_1(x_1)}{\beta_t(x_0,x_1)}\right)\beta_t(x_0,x_1)\,\pi(dx_1|x_0)\,\nu(dx_1).$$

Remark 17.27. The displacement convexity condition in Definition 17.24 becomes more stringent as β increases.

The next result is an alternative to Theorem 17.15; in many problems it is easier and more efficient to use. Recall Convention 17.10.

Theorem 17.28 (distorted displacement convexity from curvature-dimension bounds). *Let M be a Riemannian manifold, equipped with a reference measure $\nu = e^{-V}\mathrm{vol}$, where $V \in C^2(M)$. Let $K \in \mathbb{R}$ and $N \in (1, +\infty]$; let $\beta_t^{(K,N)}(x,y)$ be defined as in (14.60). Let further $p \in [2,+\infty)\cup\{c\}$ satisfy the assumptions of Theorem 17.23. Then the following three conditions are equivalent:*

(i) M satisfies the curvature-dimension bound $\mathrm{CD}(K,N)$;

(ii) Each $U \in \mathcal{DC}_N$ is displacement convex on $P_p^{\mathrm{ac}}(M)$ with distortion $(\beta_t^{(K,N)})$;

(iii) U_N is locally displacement convex with distortion $(\beta_t^{(K,N)})$.

Before explaining the proof of this result, let me state two open problems which are very natural (I have no idea how difficult they are).

Open Problem 17.29. *Is there a natural "Eulerian" counterpart to Theorem 17.28?*

Open Problem 17.30. *Theorem 17.15 and 17.28 yield two different upper bounds for $U_\nu(\mu_t)$: on one hand,*

$$U_\nu(\mu_t) \le (1-t)\,U_\nu(\mu_0) + t\,U_\nu(\mu_1) - K_{N,U}\int_0^1 \left(\int \rho_s(x)^{1-\frac{1}{N}}\,|\nabla\psi_s|^2\,d\nu\right) G(s,t)\,ds; \quad (17.25)$$

on the other hand,

$$U_\nu(\mu_t) \le (1-t)\int_M U\left(\frac{\rho_0(x_0)}{\beta_{1-t}^{(K,N)}(x_0,x_1)}\right)\beta_{1-t}^{(K,N)}(x_0,x_1)\,\pi(dx_1|x_0)\,d\nu(x_0) + t\int_M U\left(\frac{\rho_1(x_1)}{\beta_t^{(K,N)}(x_0,x_1)}\right)\beta_t^{(K,N)}(x_0,x_1)\,\pi(dx_0|x_1)\,d\nu(x_1). \quad (17.26)$$

Can one compare those two bounds, and if yes, which one is sharpest? At least in the case $N = \infty$, the second inequality implies the first one: see Theorem 30.8 at the end of these notes.

Exercise 17.31. Show, at least formally, that inequalities (17.25) and (17.26) coincide asymptotically when μ_0 and μ_1 approach each other.

Proof of Theorem 17.28. The proof shares many common points with the proof of Theorem 14.8. I shall restrict to the case $N < \infty$, since the case $N = \infty$ is very similar.

Let us start with the implication (i) $\Rightarrow$ (ii). In a first step, μ_0 and μ_1 are assumed to be compactly supported. Start again from

$$\int_M U(\rho_t(x))\, d\nu(x) = \int_M u(\delta_{t_0}(t,x))\, d\mu_{t_0}(x).$$

By applying inequality (14.55) in Theorem 14.12 (up to a factor which only depends on x and t_0, $\mathcal{D}(t)$ coincides with $\delta_{t_0}(t,x)$), and using the decreasing property of u, we get, with the same notation as in Theorem 14.12,

$$\int_M U(\rho_t(x))\, d\nu(x) \le \int_M u\Big(\tau_{K,N}^{(1-t)} \delta_{t_0}(0,x) + \tau_{K,N}^{(t)} \delta_{t_0}(1,x)\Big)\, d\mu_{t_0}(x).$$

Next, by the convexity of u, with coefficients t and $1-t$,

$$\int_M u\Big(\tau_{K,N}^{(1-t)}\, \delta_{t_0}(0,x) + \tau_{K,N}^{(t)}\, \delta_{t_0}(1,x)\Big)\, d\nu(x) \le (1-t) \int_M u\left(\frac{\tau_{K,N}^{(1-t)}}{1-t}\, \delta_{t_0}(0,x)\right)\, d\mu_{t_0}(x)$$
$$+ t \int_M u\left(\frac{\tau_{K,N}^{(t)}}{t}\, \delta_{t_0}(1,x)\right)\, d\mu_{t_0}(x).$$

Since $\beta_t^{(K,N)} = (\tau_{K,N}^{(t)}/t)^N$, the right-hand side can be rewritten as

$$(1-t) \int_M \frac{\beta_{1-t}^{(K,N)}(x_0,x_1)}{\rho_0(x_0)}\, U\left(\frac{\rho_0(x_0)}{\beta_{1-t}^{(K,N)}(x_0,x_1)}\right)\, d\pi(x_0\, x_1)$$
$$+\ t \int_M \frac{\beta_t^{(K,N)}(x_0,x_1)}{\rho_1(x_1)}\, U\left(\frac{\rho_1(x_1)}{\beta_t^{(K,N)}(x_0,x_1)}\right)\, d\pi(x_0\, x_1),$$

which is the same as the right-hand side of (17.24).

In a second step, I shall relax the assumption of compact support by a restriction argument. Let μ_0 and μ_1 be two probability measures in $P_p^{\rm ac}(M)$, and let $(Z_\ell)_{\ell\in\mathbb{N}}$, $(\mu_{t,\ell})_{0\le t\le 1,\, \ell\in\mathbb{N}}$, $(\pi_\ell)_{\ell\in\mathbb{N}}$ be as in Proposition 13.2. Let $t \in [0,1]$ be fixed. By the first step, applied with the probability measures $\mu_{t,\ell}$ and the nonlinearity $U_\ell : r \to U(Z_\ell\, r)$,

$$(U_\ell)_\nu(\mu_{t,\ell}) \le (1-t)\, (U_\ell)_{\pi_\ell,\nu}^{\beta_{1-t}^{(K,N)}}(\mu_{0,\ell})\ +\ t\, (U_\ell)\breve{\pi}_\ell, \nu^{\beta_t^{(K,N)}}(\mu_{1,\ell}). \tag{17.27}$$

It remains to pass to the limit in (17.27) as $\ell \to \infty$. The left-hand side is handled in exactly the same way as in the proof of Theorem 17.15, and the problem is to pass to the limit in the right-hand side. To ease notation, I shall write $\beta_t^{(K,N)} = \beta$. Let us prove for instance that

$$(U_\ell)_{\pi_\ell,\nu}^{\beta}(\mu_{0,\ell}) \xrightarrow[\ell\to\infty]{} U_{\pi,\nu}^{\beta}(\mu_0). \tag{17.28}$$

Since μ_0 is absolutely continuous, the optimal transport plan π comes from a deterministic transport T, and similarly the optimal transport π_ℓ comes from a deterministic transport T_ℓ; Proposition 13.2 guarantees that $T_\ell = T$, $\mu_{0,\ell}$-almost surely. So the left-hand side of (17.28) can be rewritten as

$$\int U\left(\frac{Z_\ell\, \rho_{0,\ell}(x_0)}{\beta(x_0, T(x_0))}\right)\, \beta(x_0, T(x_0))\, \nu(dx_0).$$

Since U_+ is a nondecreasing function and $Z_\ell\, \rho_{0,\ell}$ is a nondecreasing sequence, the contribution of the positive part U_+ is nondecreasing in ℓ. On the other hand, the contribution of the negative part can be controlled as in the proof of Theorem 17.23:

$$U_-\left(\frac{Z_\ell\,\rho_{0,\ell}(x_0)}{\beta(x_0,T(x_0))}\right) \le A\Big(Z_\ell\,\rho_{0,\ell}(x_0) + \beta(x_0,T(x_0))^{\frac{1}{N}}\,Z_\ell^{1-\frac{1}{N}}\,\rho_{0,\ell}(x_0)^{1-\frac{1}{N}}\Big)$$
$$\le A\big(\rho_0(x_0) + \beta(x_0,T(x_0))^{\frac{1}{N}}\rho_0(x_0)^{1-\frac{1}{N}}\big).$$

Then Theorem 17.23 and Remark 17.24 show that the latter quantity is always integrable.

As a conclusion,

$$\int U_+\left(\frac{Z_\ell\,\rho_{0,\ell}(x_0)}{\beta(x_0,T(x_0))}\right)\,\beta(x_0,T(x_0))\,\nu(dx_0) \;\xrightarrow[\ell\to\infty]{}\; \int U_+\left(\frac{\rho_0(x_0)}{\beta(x_0,T(x_0))}\right)\,\beta(x_0,T(x_0))\,\nu(dx_0)$$
by monotone convergence;

$$\int U_-\left(\frac{Z_\ell\,\rho_{0,\ell}(x_0)}{\beta(x_0,T(x_0))}\right)\,\beta(x_0,T(x_0))\,\nu(dx_0) \;\xrightarrow[\ell\to\infty]{}\; \int U_-\left(\frac{\rho_0(x_0)}{\beta(x_0,T(x_0))}\right)\,\beta(x_0,T(x_0))\,\nu(dx_0)$$
by dominated convergence.

So the limit in (17.28) holds true, and we can pass to the limit in all the terms of (17.27). This concludes the proof of (i) ⇒ (ii).

It is obvious that (ii) ⇒ (iii). So let us now consider the implication (iii) ⇒ (i). Let $x_0 \in M$, $v_0 \in T_{x_0}M$, the goal is to show that $\mathrm{Ric}_{N,\nu}(v_0) \ge K$. Construct $\widetilde{\psi}$ and ψ as in the proof of Theorem 14.8. Recall (17.19): as $\theta \to 0$,

$$H_{N,\nu}(\mu_t) - (1-t)\,H_{N,\nu}(\mu_0) - t\,H_{N,\nu}(\mu_1)$$
$$\ge -\theta^2\big(\mathrm{Ric}_{N,\nu}(v_0) + O(\theta)\big)\int_0^1\int_M \rho_s(y)^{1-\frac{1}{N}}\,d\nu(y)\,G(s,t)\,ds. \quad (17.29)$$

The change of variables $x \to T_s(x)$ is smooth and has Jacobian $\mathcal{J}_{0\to s}(x) = 1 + O(\theta)$. So

$$\int \rho_s(x)^{1-\frac{1}{N}}\,\nu(dx) = \int \rho_s(T_{0\to s}(x))^{1-\frac{1}{N}}\,\mathcal{J}_{0\to s}(x)\,\nu(dx) = \int \frac{\rho_0(x)^{1-\frac{1}{N}}}{\mathcal{J}_{0\to s}(x)^{1-\frac{1}{N}}}\mathcal{J}_{0\to s}(x)\,\nu(dx)$$
$$= \int \rho_0(x)^{1-\frac{1}{N}}\,\mathcal{J}_{0\to s}(x)^{\frac{1}{N}} = \big(1+O(\theta)\big)\left(\int \rho_0^{1-\frac{1}{N}}\,d\nu\right);$$

so (17.29) can be recast as

$$H_{N,\nu}(\mu_t) - (1-t)\,H_{N,\nu}(\mu_0) - t\,H_{N,\nu}(\mu_1)$$
$$\ge -\theta^2\,\mathrm{Ric}_{N,\nu}(v_0)\left(\frac{t(1-t)}{2}\right)\left(\int_M \rho_0^{1-\frac{1}{N}}\,d\nu\right) + O(\theta^3). \quad (17.30)$$

(Recall that $\int G(s,t)\,ds = t(1-t)/2$.)

On the other hand, by assumption the left-hand side of 17.30 is bounded above by (with obvious notation)

$$(1-t)\left(H^{\beta^{(K,N)}_{1-t}}_{N,\pi,\nu}(\mu_0) - H_{N,\nu}(\mu_0)\right) + t\left(H^{\beta^{(K,N)}_{t}}_{N,\check{\pi},\nu}(\mu_1) - H_{N,\nu}(\mu_1)\right). \quad (17.31)$$

Let us see how this expression behaves in the limit $\theta \to 0$; for instance I shall focus on the first term in (17.31). From the definitions,

$$H_{N,\pi,\nu}^{\beta_{1-t}^{(K,N)}}(\mu_0) \;-\; H_{N,\nu}(\mu_0)$$
$$= N \int \rho_0(x)^{1-\frac{1}{N}} \left(1 - \beta_{1-t}^{(K,N)}(x, T(x))^{\frac{1}{N}}\right) d\nu(x), \quad (17.32)$$

where $T = \exp(\nabla\psi)$ is the optimal transport from μ_0 to μ_1. A standard Taylor expansion shows that

$$\beta_{1-t}^{(K,N)}(x,y)^{\frac{1}{N}} = 1 + \frac{K\,[1-(1-t)^2]}{6N}\, d(x,y)^2 + O(d(x,y)^4);$$

plugging this back in (17.32), we find

$$\begin{aligned} H_{N,\pi,\nu}^{\beta_{1-t}^{(K,N)}}(\mu_0) \;-\; H_{N,\nu}(\mu_0) \\ &= -\,\frac{K\,[1-(1-t)^2]}{6} \int \rho_0(x)^{1-\frac{1}{N}} \left(\theta^2 |v_0|^2 + O(\theta^3)\right) \\ &= -\left(\theta^2 |v_0|^2 + O(\theta^3)\right) \frac{K\,[1-(1-t)^]2}{6} \left(\int \rho_0^{\,1-\frac{1}{N}}\right) d\nu. \end{aligned}$$

A similar computation can be performed for the second term in (17.31), taking into account $\int \rho_1^{1-\frac{1}{N}}\, d\nu = \int \rho^{1-\frac{1}{N}}\, d\nu + O(\theta)$. Then the whole expression (17.31) is equal to

$$-\,\theta^2\, K \left(\frac{(1-t)[1-(1-t)^2] + t[1-t^2]}{6}\right) |v_0|^2 \left(\int \rho^{1-\frac{1}{N}}\, d\nu\right) \;+\; O(\theta^3)$$
$$= -\theta^2\, \frac{K\, t(1-t)}{2}\, |v_0|^2 \left(\int \rho^{1-\frac{1}{N}}\, d\nu\right) \;+\; O(\theta^3).$$

Since this is an upper bound for the right-hand side of (17.30), we obtain after simplification

$$\mathrm{Ric}_{N,\nu}(v_0) + O(\theta) \geq |v_0|^2 + O(\theta),$$

and the conclusion follows upon taking the limit $\theta \to 0$. □

The case $N = 1$ was not addressed in Theorem 17.28, since $U_{1,\nu}$ has not been defined. However the rest of the theorem holds true:

Theorem 17.32 (Curvature-dimension bounds from displacement convexity, $N = 1$). *Let M be an n-dimensional Riemannian manifold, equipped with a reference measure $\nu = e^{-V}\mathrm{vol}$, where $V \in C^2(M)$. Let $K \in \mathbb{R}$; let $\beta_t^{(K,1)}(x,y)$ be defined as in (14.60). Let further $p \in [2, +\infty)$ satisfy Theorem 17.23 with $N = 1$. Then the following two conditions are equivalent:*

(i) M satisfies the curvature-dimension bound $\mathrm{CD}(K,1)$;

(ii) Each $U \in \mathcal{DC}_1$ is displacement convex on $P_p^{\mathrm{ac}}(M)$ with distortion $(\beta_t^{(K,1)})$;

and then necessarily $\nu = \mathrm{vol}$ and $K \leq 0$.

Proof of Theorem 17.32. When $K > 0$, (i) is obviously false since ν has to be equal to vol (otherwise $\mathrm{Ric}_{1,\nu}$ will take values $-\infty$); but (ii) is obviously false too since $\beta_t^{(K,1)} = +\infty$ for $0 < t < 1$. So we may assume that $K \leq 0$. Then the proof of (i) ⇒ (ii) is along the same lines as in Theorem 17.28. As for the implication (ii) ⇒ (i), note that $\mathcal{DC}_{N'} \subset \mathcal{DC}_1$ for all $N' < 1$, so M satisfies Condition (ii) in Theorem 17.28 with N replaced by N', and therefore $\mathrm{Ric}_{N',\nu} \geq K\, g$. If $N' < 2$, this forces M to be one-dimensional. Moreover, if V is not constant there is x_0 such that $\mathrm{Ric}_{N',\nu} = V'' - (V')^2/(N'-1)$ is < 0 for N' small enough. So V is constant and actually $\mathrm{Ric}_{1,\nu} = \mathrm{Ric} = 0$, a fortiori $\mathrm{Ric}_{1,\nu} \geq K$. □

I shall conclude this chapter with an "intrinsic" theorem of displacement convexity, in which the distortion coefficient β only depends on M and not on a priori given parameters K and N. Recall Definition 14.17 and Convention 17.10.

Theorem 17.33 (Intrinsic displacement convexity). *Let M be a Riemannian manifold with dimension n, and let $\beta_t(x, y)$ be a continuous positive function on $[0, 1] \times M \times M$. Let $p \in [2, +\infty) \cup \{c\}$ be such that the assumptions of Theorem 17.23 are satisfied with $\mathcal{X} = M$, $N = n$, $\nu = \mathrm{vol}$ and $\beta = \beta_t$ for any t. Then the following two statements are equivalent:*

(a) $\beta \le \overline{\beta}$;

(b) For all $U \in \mathcal{DC}_n$, the functional U_ν is displacement convex on $P_p^{\mathrm{ac}}(M)$ with distortion coefficients β.

The proof of this theorem follows the same lines as the proof of Theorem 17.28, with the help of Theorem 14.20; details are left to the reader.

Bibliographical Notes

The definition of displacement convexity classes goes back to McCann's PhD [265], in the case $N < \infty$. (McCann required u in Definition 17.1 to be nonincreasing, but this is automatic as noticed in Remark 17.2.) The definition of $\mathcal{DC}_\infty$ is taken from [247]. Conditions (i), (ii) or (iii) in Definition **??** occur in various contexts, in particular in the theory of nonlinear diffusion equations (as we shall see in Chapter 23), so it is normal that these classes of nonlinearities were rediscovered later by several authors. The normalization $U(0) = 0$ is not the only possible one, but it has many advantages. In [265] or more recently [247] it is not imposed that U should be twice differentiable on $(0, \infty)$.

Many authors have contributed to Theorem 17.15 and the story is roughly as follows.

McCann [267] proved the displacement convexity of U_ν when $M = \mathbb{R}^n$, $n = N$ and ν is the Lebesgue measure. Things were made quite simpler by the Euclidean setting (no Jacobi fields, no $d^2/2$-convex functions, etc.) and by the fact that only displacement convexity (as opposed to Λ-displacement convexity) was considered. Apart from that, the strategy was essentially the same as the one used in this chapter, based on a change of variables, except that the reference measure was μ_0 instead of μ_{t_0}. McCann's proof was recast in my book [365, Proof of Theorem 5.15 (i)]; it takes only a few lines, once one has accepted (a) the *concavity of* $\det^{1/n}$ in $\mathbb{R}^n$: that is, if a symmetric matrix $S \le I_n$ is given, then $t \longmapsto \det(I_n - tS)^{1/n}$ is concave [365, Lemma 5.21]; and (b) the change of variables formula along displacement interpolation.

Later Cordero-Erausquin, McCann and Schmuckenschläger [118] studied genuinely Riemannian situations, replacing the concavity of $\det^{1/n}$ in $\mathbb{R}^n$ by distortion estimates, and extending the formula of change of variables along displacement interpolation. With these tools they basically proved the displacement convexity of U_ν for $U \in \mathcal{DC}_N$, as soon as M is a Riemannian manifold of dimension $n \le N$ and nonnegative Ricci curvature, with the reference measure $\nu = \mathrm{vol}$. It is clear from their paper that their argument also yields, for instance, K-displacement convexity of H as soon as $\mathrm{Ric} \ge K$; moreover, they established Theorem 17.33 for compactly supported densities.

Several authors independently felt the need to rewrite more explicitly the connection between Jacobi fields and optimal transport, that was implicit in [118]. This was done simultaneously by Cordero-Erausquin, McCann and Schmuckenschläger [117] again; by Sturm [337]; and by Lott and myself [247]. All those arguments heavily draw on [118]; they

are also reminiscent of arguments used in the proof of the Lévy–Gromov isoperimetric inequality. A large part of the proofs were actually devoted to establish the Jacobian estimates on the exponential function, which I recast here as part of Chapter 14.

Modifications needed to replace the volume measure by $\nu = e^{-V}\mathrm{vol}$ were discussed by Sturm [337] for $N = \infty$; and independently by Lott and myself [247] for $N \leq \infty$. For the purpose of these notes, all those modifications were included in the section about "change of measure" in Chapter 14.

It was first proven by Sturm and Von Renesse [341] that the displacement convexity of H does not only result from, but actually characterizes the nonnegativity of the Ricci curvature. This statement was generalized by Lott and the author [247], and independently Sturm [335].

Theorem 17.28 is due to Sturm [339] in the most important case $U = U_N$. Then the general formulation with arbitrary $U \in \mathcal{DC}_N$ was worked out shortly after by Lott and myself [249]. The proof rests on the inequality (14.55) in Theorem 14.12, which is (as far as I know) due to Cordero-Erausquin, McCann and Schmuckenschläger [118]. All this was for $N < \infty$; then the case $N = \infty$ works the same, once one has the correct definitions for $\mathcal{DC}_\infty$ and $\beta_t^{(K,\infty)}$.

The use of Theorem 17.8 to control noncompactly supported probability densities is essentially taken from Lott and myself [247]; the only change with respect to that reference is that I do not try to define U_ν on the whole of P_2^{ac}, and therefore do not require p to be equal to 2.

In this chapter I used restriction arguments to remove the compactness assumption. An alternative strategy consists in using a density argument and stability theorems (as in [247]); these tools will be examined in Part III. In the particular case when the manifold has nonnegative sectional curvature, it is also possible to directly apply the argument of change of variables to the family (μ_t), even if it is not compactly supported, thanks to the uniform inequality (8.41).

Another innovation in the proofs of this chapter is the idea of choosing μ_{t_0} as the reference measure with respect to which changes of variables are performed. The advantage of that procedure (which evolved from discussions with Ambrosio) is that the transport map from μ_{t_0} to μ_t is Lipschitz for all times t, as we know from Chapter 8; while the transport map from μ_0 to μ_1 is only of bounded variation. So the proof given in this section only uses the Jacobian formula for *Lipschitz* changes of variables, and not the more subtle formula for BV changes of variables.

Paths $(\mu_t)_{0\leq t\leq 1}$ defined in terms of transport from a given measure $\widetilde{\mu}$ (not necessarily of the form μ_{t_0}) are studied in [15] in the context of generalized geodesics in $P_2(\mathbb{R}^n)$. The procedure amounts to consider $\mu_t = (T_t)_\# \widetilde{\mu}$ with $T_t(x) = (1-t)\,T_0(x) + t\,T_1(x)$, where T_0 is optimal between $\widetilde{\mu}$ and μ_0, and T_1 is optimal between $\widetilde{\mu}$ and μ_1. Displacement convexity theorems work for these generalized geodesics just as well as for the true geodesics, and they are useful in error estimates for gradient flows. It is not clear whether there is a Riemannian analogue.

The proofs in the present chapter are of Lagrangian nature, but, as I said before, it is also possible to go for an Eulerian proof, at the price of further regularization procedures (that are messy but more or less standard), see in particular Otto and Westdickenberg [294]. As pointed out by Otto, the Eulerian point of view, although more technical, has the merit to separate very clearly the input from local smooth differential geometry (Bochner's formula is a purely local statement about the Laplace operator on M, seen as a differential operator on very smooth functions) and the input from global nonsmooth

analysis (Wasserstein geodesics involve $d^2/2$-convexity, which is a nonlocal condition; and $d^2/2$-convex functions are in general nonsmooth).

Apart from functionals of the form U_ν, all interesting examples of displacement convex functionals presently known (at least to me) are constructed with functionals of the form $\Phi : \mu \longmapsto \int \Phi(x)\, d\mu(x)$, or $\Psi : \mu \longmapsto \int \Psi(x,y)\, d\mu(x)\, d\mu(y)$, where Φ is a given "potential" and Ψ is a given "interaction potential" [103, 102, 39]. It is easy to show that the displacement convexity of Φ (seen as a function on $P_2(M)$) is equivalent to the geodesic convexity of Φ, seen as a function on M. Similarly, it is not difficult to show that the displacement convexity of Ψ is equivalent to the geodesic convexity of Ψ, seen as a function on $M \times M$. These results can be found for instance in my book [365, Theorem 5.15] in the Euclidean setting. There it is assumed there that $\Psi(x,y) = \Psi(x-y)$, with Ψ convex, but it is immediate to generalize the proof to the case where Ψ is convex on $\mathbb{R}^n \times \mathbb{R}^n$.

There is no interesting displacement convexity statement known for the Coulomb interaction potential; however, Blower [54] proved that

$$E(\mu) = \frac{1}{2} \int_{\mathbb{R}^2} \log \frac{1}{|x-y|}\, \mu(dx)\, \mu(dy)$$

defines a displacement convex functional on $P_2^{\mathrm{ac}}(\mathbb{R})$. Blower also studied what happens when one adds a potential energy to E, and used these tools to establish concentration inequalities for the eigenvalues of some large random matrices.

Exercise 17.34. Prove the statement alluded to above: If M is a compact Riemannian manifold and Ψ a function on $M \times M$, then Ψ defines a displacement functional on $P_2(M)$ if and only if it is geodesically convex on $M \times M$. Hint: A product of geodesics in M is also a geodesic in $M \times M$.

18

Volume control

Controlling the volume of balls is a universal problem in geometry. This means of course controlling the volume from above when the radius increases to infinity; but also controlling the volume from below when the radius decreases to 0. The *doubling property* is useful in both situations.

Definition 18.1 (doubling property). *Let $(\mathcal{X}, d)$ be a metric space, and let μ be a Borel measure on $\mathcal{X}$, not identically 0. The measure μ is said to be doubling if there exists a constant D such that*

$$\forall x \in \mathcal{X}, \quad \forall r > 0, \quad \nu[B_{2r}(x)] \leq D\, \nu[B_r(x)]. \tag{18.1}$$

The measure μ is said to be locally doubling if for any fixed ball $B_R(z) \subset \mathcal{X}$, there is a constant $D = D(z, R)$ such that

$$\forall x \in B_R(z), \quad \forall r > 0, \quad \nu[B_{2r}(x)] \leq D\, \nu[B_r(x)]. \tag{18.2}$$

Remark 18.2. It is equivalent to say that a measure ν is locally doubling, or that its restriction to any ball $B(z, R)$ (considered as a metric space) is doubling.

When the distance d and the reference measure ν are clear from the context, I shall often say that the space $\mathcal{X}$ is doubling (resp. locally doubling), instead of writing that the measure ν is doubling on the metric space $(\mathcal{X}, d)$.

It is a standard fact in Riemannian geometry that doubling constants may be estimated, at least locally, in terms of curvature-dimension bounds. These estimates express the fact that the manifold does not contain *sharp spines*. Of course, this is obvious for a Riemannian manifold, since it is locally diffeomorphic to an open subset of $\mathbb{R}^n$; but curvature-dimension bounds quantify this in terms of the intrinsic geometry, without reference to charts.

Another property which is obvious for a smooth Riemannian manifold, but which doubling makes quantitative, is the fact that the reference measure has full support:

Proposition 18.3. *Let $(\mathcal{X}, d)$ be a metric space equipped with a locally doubling measure ν. Then* $\mathrm{Spt}\, \nu = \mathcal{X}$.

Proof. Let $x \in \mathcal{X}$, and let $r > 0$. Since ν is nonzero, there is $R > 0$ such that $\nu[B_R(x)] > 0$. Then there is a constant C, possibly depending on x and R, such that ν is C-doubling inside $B_R(x)$. Let $n \in \mathbb{N}$ be large enough that $R \leq 2^n r$; then

$$0 < \nu[B_R(x)] \leq C^n \nu[B_r(x)].$$

So $\nu[B_r(x)] > 0$. Since r is arbitrarily small, x has to lie in the support of ν. □

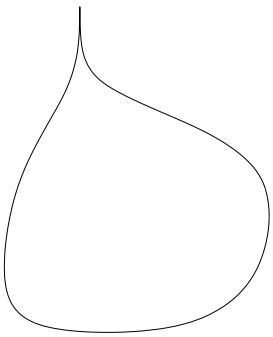


Fig. 18.1. The natural volume measure on this "singular surface" (a balloon with a spine) is not doubling.

One of the goals of this chapter is to get doubling constants from curvature-dimension bounds, by means of arguments based on optimal transport. This is not the standard strategy, but it will work just as well as any other, since the results in the end will be optimal. As a preliminary step, I shall establish a "distorted" version of the famous **Brunn–Minkowski inequality**.

Distorted Brunn–Minkowski inequality

The classical Brunn–Minkowski inequality states that whenever A_0 and A_1 are two nonempty compact subsets of $\mathbb{R}^n$, then

$$\big|A_0 \,+\, A_1\big|^{\frac{1}{n}} \geq |A_0|^{\frac{1}{n}} + |A_1|^{\frac{1}{n}}, \tag{18.3}$$

where $|\cdot|$ stands for Lebesgue measure, and $A_0 + A_1$ is the set of all vectors of the form $a_0 + a_1$ with $a_0 \in A_0$ and $a_1 \in A_1$. This inequality contains the Euclidean isoperimetric inequality as a limit case (take $A_1 = \varepsilon B(0,1)$ and let $\varepsilon \to 0$).

It is not obvious to guess the "correct" generalization of (18.3) to general Riemannian manifolds, and it is only a few years ago that a plausible answer to that problem emerged, in terms of the distortion coefficients (14.60).

In the sequel, I shall use the following notation: if A_0 and A_1 are two nonempty compact subsets of a Riemannian manifold M, then $[A_0, A_1]_t$ stands for the set of all *t-barycenters* of A_0 and A_1, that is the set of all $y \in M$ that can be written as γ_t, where γ is a minimizing, constant-speed geodesic with $\gamma_0 \in A_0$ and $\gamma_1 \in A_1$. Equivalently, $[A_0, A_1]_t$ is the set of all y such that there exists $(x_0, x_1) \in A_0 \times A_1$ with $d(x_0, y)/d(y, x_1) = t/(1-t)$.

Theorem 18.4 (Distorted Brunn–Minkowski inequality). *Let M be a Riemannian manifold equipped with a reference measure $\nu = e^{-V}\,\mathrm{vol}$, $V \in C^2(M)$, satisfying a curvature-dimension condition $\mathrm{CD}(K,N)$. Let A_0, A_1 be two nonempty compact subsets, and let $t \in (0,1)$. Then*

- If $N < \infty$,

$$\nu\big[[A_0,A_1]_t\big]^{\frac{1}{N}} \geq (1-t)\left[\inf_{(x_0,x_1)\in A_0\times A_1} \beta_{1-t}^{(K,N)}(x_0,x_1)^{\frac{1}{N}}\right] \nu[A_0]^{\frac{1}{N}} + \,t\left[\inf_{(x_0,x_1)\in A_0\times A_1} \beta_t^{(K,N)}(x_0,x_1)^{\frac{1}{N}}\right] \nu[A_1]^{\frac{1}{N}}, \tag{18.4}$$

where $\beta_t^{(K,N)}(x_0,x_1)$ are the distortion coefficients defined in (20.12).

- *If $N = \infty$, then*

$$\log \frac{1}{\nu\big[[A_0, A_1]_t\big]} \le (1-t)\, \log \frac{1}{\nu[A_0]} \;+\; t\, \log \frac{1}{\nu[A_1]} \;-\; \frac{K\, t(1-t)}{2} \sup_{x_0 \in A_0,\, x_1 \in A_1} d(x_0, x_1)^2.$$

By particularizing Theorem 18.4 to the case when $K = 0$ and $N < \infty$ (so $\beta_t^{(K,N)} = 1$), one can show that nonnegatively curved Riemannian manifolds satisfy a Brunn–Minkowski inequality which is formally very similar to the Brunn–Minkowski inequality in $\mathbb{R}^n$:

Corollary 18.5 (Brunn–Minkowski inequality for nonnegatively curved manifolds). *With the same notation as in Theorem 18.4, if M satisfies the curvature-dimension condition* $\mathrm{CD}(0, N)$*, $N \in (1, +\infty)$, then*

$$\nu\big[[A_0, A_1]_t\big]^{\frac{1}{N}} \ge (1-t)\, \nu[A_0]^{\frac{1}{N}} \;+\; t\, \nu[A_1]^{\frac{1}{N}}. \tag{18.5}$$

Remark 18.6. When $M = \mathbb{R}^n$, $N = n$, inequality (18.5) reduces to

$$\big|(1-t)A_0 + tA_1\big|^{\frac{1}{n}} \ge (1-t)|A_0|^{\frac{1}{n}} + t|A_1|^{\frac{1}{n}},$$

where $|\ \cdot\ |$ stands for the n-dimensional Lebesgue measure. By homogeneity, this is equivalent to (18.4).

Idea of the proof of Theorem 18.4. Introduce an optimal coupling between a random point γ_0 chosen uniformly in A_0 and a random point γ_1 chosen uniformly in A_1 (as in the proof of isoperimetry in Chapter 2). Then γ_t is a random point (not necessarily uniform) in A_t. If A_t would be very small, then the law μ_t of γ_t would be very concentrated, so its density would be very high, but then this would contradict the displacement convexity estimates implied by the curvature assumptions. For instance, consider for simplicity $U(r) = r^m$, $m \ge 1$, $K = 0$: since $U_\nu(\mu_0)$ and $U_\nu(\mu_1)$ are finite, this implies a bound on $U_\nu(\mu_t)$, and this bound cannot hold if the support of μ_t is too small (in the extreme case where A_t would be a single point, μ_t would be a Dirac mass and $U_\nu(\mu_t)$ should be $+\infty$). It turns out that the optimal estimates are obtained with $U = U_N$, as defined in (16.15). □

Detailed proof of Theorem 18.4. First consider the case $N < \infty$. For brevity I shall write just β_t instead of $\beta_t^{(K,N)}$. By regularity of the measure ν and an easy approximation argument, it is sufficient to treat the case when $\nu[A_0] > 0$ and $\nu[A_1] > 0$. Then one may define $\mu_0 = \rho_0\, \nu$, $\mu_1 = \rho_1\, \nu$, where

$$\rho_0 = \frac{1_{A_0}}{\nu[A_0]}, \qquad \rho_1 = \frac{1_{A_1}}{\nu[A_1]}.$$

In words, μ_{t_0} ($t_0 \in \{0, 1\}$) is the law of a random point distributed uniformly in A_i. Let $(\mu_t)_{0 \le t \le 1}$ be the unique displacement interpolation between μ_0 and μ_1, for the cost function $d(x, y)^2$. Since M satisfies the curvature-dimension bound $\mathrm{CD}(K, N)$, Theorem 17.28 and Lemma 29.4, applied with $U(r) = U_N(r) = -N\big(r^{1-\frac{1}{N}} - r\big)$, imply

$$\begin{aligned}\int_M U_N(\rho_t(x))\, \nu(dx) \le\ & (1-t) \int_M U_N\left(\frac{\rho_0(x_0)}{\beta_{1-t}(x_0, x_1)}\right) \frac{\beta_{1-t}(x_0, x_1)}{\rho_0(x_0)}\, \pi(dx_0\, dx_1) \\ & +\ t \int_M U_N\left(\frac{\rho_1(x_1)}{\beta_t(x_0, x_1)}\right) \frac{\beta_t(x_0, x_1)}{\rho_1(x_1)}\, \pi(dx_0\, dx_1), \end{aligned} \tag{18.6}$$

where π is the optimal coupling of (μ_0, μ_1), and $\beta_t(x_0, x_1)$ is defined by (20.12). After replacement of U_N by its explicit expression and simplification, this leads to

$$\int_M \rho_t(x)^{1-\frac{1}{N}}\, \nu(dx) \geq (1-t) \int_M \rho_0(x)^{-\frac{1}{N}} \beta_{1-t}(x_0, x_1)^{\frac{1}{N}}\, \pi(dx_0\, dx_1) + t \int_M \rho_1(x)^{-\frac{1}{N}} \beta_t(x_0, x_1)^{\frac{1}{N}}\, \pi(dx_0\, dx_1). \quad (18.7)$$

Since π is supported in $A_0 \times A_1$ and has marginals $\rho_0\, \nu$ and $\rho_1 \nu$, one can bound the right-hand side below by

$$(1-t)\, \beta_{1-t}^{\frac{1}{N}} \int_M \rho_0(x_0)^{1-\frac{1}{N}}\, d\nu(x_0) \;+\; t\, \beta_t^{\frac{1}{N}} \int_M \rho_1(x_1)^{1-\frac{1}{N}}\, d\nu(x_1),$$

where β_t stands for the minimum of $\beta_t(x_0, x_1)$ over all pairs $(x_0, x_1) \in A_0 \times A_1$. Then, by explicit computation,

$$\int_M \rho_0(x_0)^{1-\frac{1}{N}}\, d\nu(x_0) = \nu[A_0]^{\frac{1}{N}}, \qquad \int_M \rho_1(x_1)^{1-\frac{1}{N}}\, d\nu(x_1) = \nu[A_1]^{\frac{1}{N}}.$$

So to conclude the proof of (18.4) it sufficient to show

$$\int_M \rho_t^{1-\frac{1}{N}}\, d\nu \leq \nu\big[[A_0, A_1]_t\big]^{\frac{1}{N}}.$$

Obviously, μ_t is supported in $A_t = [A_0, A_1]_t$; therefore ρ_t is a probability density on that set. By Jensen's inequality,

$$\begin{aligned} \int_{A_t} \rho_t^{1-\frac{1}{N}}\, d\nu &= \nu[A_t] \int_{A_t} \rho_t^{1-\frac{1}{N}}\, \frac{d\nu}{\nu[A_t]} \\ &\leq \nu[A_t] \left(\int_{A_t} \rho_t\, \frac{d\nu}{\nu[A_t]} \right)^{1-\frac{1}{N}} \\ &= \nu[A_t]^{\frac{1}{N}} \left(\int_{A_t} \rho_t\, d\nu \right)^{1-\frac{1}{N}} = \nu[A_t]^{\frac{1}{N}}. \end{aligned}$$

This concludes the proof of (18.4).

The proof in the case $N = \infty$ follows the same lines, except that now it is based on the K-displacement convexity of H_ν and the convexity of $r \longmapsto r \log r$. □

Bishop–Gromov inequality

The Bishop–Gromov inequality states that the volume of balls in a space satisfying $\mathrm{CD}(K, N)$ does not grow faster than the volume of balls in the model space of constant sectional curvature having Ricci curvature equal to K and dimension equal to N. In the case $K = 0$, it takes the following simple form:

$$\frac{\nu[B_r(x)]}{r^N} \qquad \text{is a nonincreasing function of } r.$$

In the cases $K > 0$ (resp. $K < 0$), the quantity on the left-hand side should be replaced by

$$\frac{\nu[B_r(x)]}{\displaystyle\int_0^r \left(\sin\sqrt{\frac{K}{N-1}}\, t\, dt\right)^{N-1} dt} \qquad \left(\quad \text{resp.} \quad \frac{\nu[B_r(x)]}{\displaystyle\int_0^r \left(\sinh\sqrt{\frac{|K|}{N-1}}\, t\, dt\right)^{N-1} dt} \qquad \right).$$

Here is a precise statement:

Theorem 18.7 (Bishop–Gromov inequality). *Let M be a Riemannian manifold equipped with a reference measure $\nu = e^{-V}\,\mathrm{vol}$, satisfying the curvature-dimension condition $\mathrm{CD}(K,N)$ for some $K \in \mathbb{R}$, $1 < N < \infty$. Let further*

$$s^{(K,N)}(t) = \begin{cases} \left(\sin\sqrt{\dfrac{K}{N-1}}\, t\right)^{N-1} & \text{if } K > 0 \\ t^{N-1} & \text{if } K = 0 \\ \left(\sinh\sqrt{\dfrac{|K|}{N-1}}\, t\right)^{N-1} & \text{if } K < 0 \end{cases}$$

Then, for any $x \in M$,

$$\frac{\nu[B_r(x)]}{\displaystyle\int_0^r s^{(K,N)}(t)\, dt}$$

is a nonincreasing function of r.

Proof of Theorem 18.7. Let us start with the case $K = 0$ which is simpler. Let $A_0 = \{x\}$ and $A_1 = B_r(x)$; in particular, $\nu[A_0] = 0$. For any $s \in (0, r)$, one has $[A_0, A_1]_{\frac{s}{r}} \subset B_s(x)$, so by the Brunn–Minkowski inequality (18.5),

$$\nu[B_s(x)]^{\frac{1}{N}} \geq \nu\big[[A_0, A_1]_{\frac{s}{r}}\big] \geq \left(\frac{s}{r}\right) \nu[B_r(x)]^{\frac{1}{N}},$$

from which the claim follows immediately.

For the general case, it is sufficient to check that

$$\frac{\frac{d}{dr}\,\nu[B_r(x)]}{s^{(K,N)}(r)} \qquad \text{is nonincreasing.} \tag{18.8}$$

Indeed, it is easy to convince oneself that if f and g are positive functions on $\mathbb{R}_+$ with f/g nonincreasing, then also $(\int_0^r f)/(\int_0^r g)$ is a nonincreasing function of r. (From our assumptions, the function $\nu[B_r(x)]$ is a differentiable function of r; but even if that were not the case it would not be so difficult to get around this problem.)

Apply Theorem 18.4 with $A_0 = \{x\}$ again, but now $A_1 = B_{r+\varepsilon}(x) \setminus B_r(x)$; then for $t \in (0,1)$ one has $[A_0, A_1]_t \subset B_{t(r+\varepsilon)}(x)$. Moreover, for $K \geq 0$, one has

$$\beta_t^{(K,N)}(x_0, x_1) \geq \left(\frac{\sin\left(t\sqrt{\frac{K}{N-1}}(r+\varepsilon)\right)}{t\sin\left(\sqrt{\frac{K}{N-1}}(r+\varepsilon)\right)}\right)^{N-1};$$

for $K < 0$ the same formula remains true with sin replaced by sinh, K by $|K|$ and $r+\varepsilon$ by $r-\varepsilon$. In the sequel, I shall only consider $K > 0$, the treatment of $K < 0$ being obviously

similar. After applying the above bounds and raising each side of inequality (18.4) to the power N, one obtains

$$\frac{\nu\big[B_{t(r+\varepsilon)}(x) \setminus B_{tr}(x)\big]}{\Big(t\,\sin\sqrt{\frac{K}{N-1}}(tr+t\varepsilon)\Big)^{N-1}} \;\geq\; t^N \frac{\nu\big[B_{r+\varepsilon}(x) \setminus B_r(x)\big]}{\Big(\sin\sqrt{\frac{K}{N-1}}(r+\varepsilon)\Big)^{N-1}}.$$

If $\phi(r)$ stands for $\nu[B_r(x)]$, then the above inequality can be rewritten as

$$\frac{\phi(tr+t\varepsilon)-\phi(tr)}{t\,s^{(K,N)}(t(r+\varepsilon))} \;\geq\; \frac{\phi(r+\varepsilon)-\phi(r)}{s^{(K,N)}(r)}.$$

In the limit $\varepsilon \to 0$, this yields

$$\frac{\phi'(tr)}{s^{(K,N)}(tr)} \geq \frac{\phi'(r)}{s^{(K,N)}(r)},$$

so $\phi'/s^{(K,N)}$ is indeed nonincreasing. □

Doubling property

From Theorem 18.7 and elementary estimates on the function $s^{(K,N)}$ it is easy to deduce the following corollary:

Corollary 18.8 (doubling property). *Let M be a Riemannian manifold equipped with a reference measure $\nu = e^{-V}\,vol$, satisfying the curvature-dimension condition $\mathrm{CD}(K,N)$ for some $K \in \mathbb{R}$, $1 < N < \infty$. Then ν is doubling with a constant C that is*

- *uniform and no more than 2^N if $K \geq 0$;*
- *locally uniform and no more than $2^N\,D(K,N,R)$ if $K < 0$, where*

$$D(K,N,R) = \left[\cosh\left(2\sqrt{\frac{|K|}{N-1}}\,R\right)\right]^{N-1}, \tag{18.9}$$

when restricted to a large ball $B(z,R)$.

The Bishop–Gromov inequality is however more precise than just doubling property: for instance, if $0 < s < r$ then, with the same notation as before,

$$\nu[B_r(x)] \geq \nu[B_s(x)] \geq \left(\frac{V(s)}{V(r)}\right)\,\nu[B_r(x)],$$

where $V(r)$ is the volume of $B_r(x)$ in the model space. It follows that $\nu[B_r(x)]$ is a continuous function of r. Of course, this property is otherwise obvious, but the Bishop–Gromov inequality provides an explicit modulus of continuity.

There do not seem to be any "natural" analogues of these results in the case $N = \infty$.

Bibliographical Notes

The Brunn–Minkowski inequality in $\mathbb{R}^n$ goes back to the end of the nineteenth century; it was first established by Brunn (for convex sets in dimension 2 or 3), and later generalized by Minkowski (for convex sets in arbitrary dimension) and Lusternik (for arbitrary compact sets). Nowadays, it is still one of the cornerstones of the geometry of convex bodies. Standard references on the Brunn–Minkowski theory are the book by Schneider [329] and the more recent survey paper by Gardner [181]; see also Maurey's lecture [263].

It is classical to prove the Brunn–Minkowski (in $\mathbb{R}^n$) via changes of variables, usually called *reparametrizations* in this context. McCann [265] noticed that optimal transport could be used to yield a convenient reparametrization; this is a bit more complicated than the reparametrizations classically used in $\mathbb{R}^n$, but it has the advantage to be defined in more intrinsic terms. McCann's argument is reproduced in [365, Section 6.1]; it is basically the same as the proof of Theorem 18.4, only much simpler because it is in Euclidean space.

At the end of the nineties, it was still not clear what would be the correct extension of that theory to curved spaces. The first hint came when Cordero-Erausquin [114] used the formalism of optimal transport to guess a Prékopa–Leindler inequality on the sphere. In Euclidean space, the Prékopa–Leindler inequality is a well-known functional version of the Brunn–Minkowski inequality (it is discussed for instance in the above-mentioned surveys, and we shall meet it in the next chapter). Then Cordero-Erausquin, McCann and Schmuckenschläger [118] developed the tools necessary to make this approach rigorous, and also established Prékopa–Leindler inequalities in curved geometry (when the reference measure is the volume). Then Sturm [339] adapted the proof of [118] to get Brunn–Minkowski inequalities, for general reference measures.

The proof of the Bishop–Gromov inequality in the case $K = 0$ is taken from [247]. Apart from that, my presentation in this chapter is strongly inspired by Sturm [339]. In particular, it is from that work that I took the statement of Theorem 18.4 and the proof of the Bishop–Gromov inequality for $K \neq 0$.

More classical proofs of the Bishop–Gromov inequality can be found in reference textbooks such as [175].

19

Density control and local regularity

The following situation occurs in many problems of local regularity: Knowing a certain estimate on a certain ball $B_r(x_0)$, deduce a better estimate on a smaller ball, say $B_{r/2}(x_0)$. In the fifties, this point of view was put to a high degree of sophistication by De Giorgi in his famous proof of Hölder estimates for solutions of elliptic second-order partial differential equations; and it also plays a role in the alternative solutions found at the same time by Nash, and later by Moser. When fine analysis on metric spaces started to develop, it became an important issue to understand what were the key ingredients lying at the core of the methods of De Giorgi, Nash and Moser. It is now accepted by many that the two key inequalities are

- a **doubling inequality** for the reference volume measure;

- a **local Poincaré inequality**, controlling the deviation of a function on a smaller ball by the integral of its gradient on a larger ball. Here below is a precise definition:

Definition 19.1 (local Poincaré inequality). *Let $(\mathcal{X}, d)$ be a metric space and let ν be a Borel measure on $\mathcal{X}$. It is said that ν satisfies a local Poincaré inequality with constant C if, for any Lipschitz function u, any point $x_0 \in \mathcal{X}$ and any radius $r > 0$,*

$$\fint_{B_r(x_0)} \Big| u(x) - \langle u \rangle_{B_r(x_0)} \Big| \, d\nu(x) \le C r \fint_{B_{2r}(x_0)} |\nabla u(x)| \, d\nu(x), \tag{19.1}$$

where $\fint_B = (\nu[B])^{-1} \int_B$ is the averaged integral over B, and $\langle u \rangle_B = \fint_B u \, d\nu$ is the average of the function u on B.

Let B be a Borel subset of $\mathcal{X}$. It is said that ν satisfies a local Poincaré inequality with constant C on B if inequality (19.1) *holds true under the additional restriction that $B_{2r}(x_0) \subset B$.*

Here the word "local" means that the inequality is interested in averages *around some point* x_0. This is in contrast with the "global" Poincaré inequalities that will be considered later in Chapter 21, in which averages are over the whole space.

There are an incredible number of variants of Poincaré inequalities, but I shall stick to the ones appearing in Definition 19.1. Sometimes I shall say that ν satisfies a *uniform* local Poincaré inequality to stress the fact that the constant C is independent of x_0 and r. For most applications this uniformity is not important, all that matters is that inequality (19.1) holds true in the neighborhood of any point x_0; so it is sufficient to prove that ν satisfies a local Poincaré inequality with constant $C = C(R)$ on each ball $B(z, R)$, where z is fixed once for all.

Just as the doubling inequality, the local Poincaré inequality might be ruined by sharp spines, and Ricci curvature bounds will prevent those spines to occur, providing quantitative Poincaré constants (that will be uniform in nonnegative curvature). Again, the goal of this chapter is to prove these facts by using optimal transport. The strategy goes through *pointwise bounds* on the density of the displacement interpolant.

There are at least two ways to prove pointwise bounds on the displacement interpolant. The first one consists in combining the Jacobian equation involving the density of the interpolant (Chapter 11) with the Jacobian estimates derived from the Ricci curvature bounds (Chapter 14). The second way goes via displacement convexity (Chapter 17); it is quite more indirect, but its interest will become apparent in the last chapter of these notes.

Of course, pointwise bounds do not result directly from displacement convexity, which only yields *integral* bounds on the interpolant; however, it is possible to deduce pointwise bounds from integral bounds by using the stability of optimal transport under *restriction* (recall Theorem 4.5). The idea is simple: a pointwise bound on $\rho_t(x)$, will be achieved by considering integral bounds on a very small ball $B_\delta(x)$, as $\delta \to 0$.

Apart from the local Poincaré inequality, the pointwise control on the density will imply at once the Brunn–Minkowski inequality, and also its functional counterpart, the Prékopa–Leindler inequality. This is not surprising, since a pointwise control is morally stronger than an integral control.

Pointwise estimates on the interpolant density

Theorem 19.2 (pointwise bounds on the displacement interpolant from curvature-dimension). *Let M be a Riemannian manifold equipped with a reference measure $\nu = e^{-V}\,\mathrm{vol}$, $V \in C^2(M)$, satisfying a curvature-dimension $\mathrm{CD}(K,N)$ for some $N \in (1,\infty]$, $K \in \mathbb{R}$. Let further $\mu_0 = \rho_0\,\nu$ and $\mu_1 = \rho_1\,\nu$ be two probability measures in $P_p^{\mathrm{ac}}(M)$, where $p \in [2,+\infty) \cup \{c\}$ satisfies the assumptions of Theorem 17.8. Let $(\mu_t)_{0\le t\le 1}$ be the unique displacement interpolation between μ_0 and μ_1, and let ρ_t stand for the density of μ_t with respect to ν. Then*

- If $N < \infty$, one has the pointwise bound

$$\rho_t(x) \le \sup_{x\in[x_0,x_1]_t} \left((1-t)\Big(\frac{\rho_0(x_0)}{\beta^{(K,N)}_{1-t}(x_0,x_1)}\Big)^{-\frac{1}{N}} \;+\; t\Big(\frac{\rho_1(x_1)}{\beta^{(K,N)}_{t}(x_0,x_1)}\Big)^{-\frac{1}{N}} \right)^{-N}, \qquad (19.2)$$

where by convention $\big((1-t)\,a^{-\frac{1}{N}} + t\,b^{-\frac{1}{N}}\big)^{-N} = 0$ if either a or b is 0;

- If $N = \infty$, one has the pointwise bound

$$\rho_t(x) \le \sup_{x\in[x_0,x_1]_t} \rho_0(x_0)^{1-t}\,\rho_1(x_1)^t\,\exp\left(-\frac{K\,t(1-t)}{2}\,d(x_0,x_1)^2\right). \qquad (19.3)$$

As I said before, there are (at least) two possible schemes of proof for Theorem 19.2. The first one is by direct application of the Jacobian estimates from Chapter 14; the second one is based on the displacement convexity estimates from Chapter 17. The first one is formally simpler, while the second one has the advantage to be based on very robust functional inequalities. I shall only sketch the first proof, forgetting about regularity issues; and give a detailed treatment of the second one.

Sketch of proof of Theorem 19.2 by Jacobian estimates. Let us pretend that Theorem 10.35 applies. Let $\nabla\psi$ be a gradient of $(d^2/2)$-convex function so that $\mu_1 = [\exp(\nabla\psi)]_\# \mu_0$; then $\mu_t = [\exp(t\nabla\psi)]_\#\mu_0$. Let $\mathcal{J}(t,x)$ stand the Jacobian determinant of $\exp(t\nabla\psi)$; then, with the shorthand $x_t = \exp_{x_0}(t\nabla\psi(x_0))$, the Jacobian equation of change of variables can be written

$$\rho_0(x_0) = \rho_t(x_t)\,\mathcal{J}(t,x_0).$$

Similarly,

$$\rho_0(x_0) = \rho_1(x_1)\,\mathcal{J}(1,x_0).$$

Then the result follows directly from Theorems 14.11 and 14.12: Apply equation (14.55) if $N < \infty$, (14.54) if $N = \infty$ (recall that $\mathcal{D} = \mathcal{J}^{\frac{1}{N}}$, $\ell = -\log \mathcal{J}$). □

Proof of Theorem 19.2 by displacement interpolation. For simplicity I shall only consider the case $N < \infty$, and derive the conclusion from Theorem 17.28. Then the case $N = \infty$ can be treated either by adapting the proof of the case $N < \infty$ (replacing Theorem 17.28 by Theorem 17.15, and using the function U_∞ defined in (16.15)), or by taking the limit $N \to \infty$ in (19.2).

Let $t \in [0,1]$ be given, $(\mu_s)_{0\le s\le 1}$ as in the statement of the theorem, and let Π be the law of a random geodesic γ such that law $(\gamma_s) = \mu_s$. Let now y be an arbitrary point in M, and $\delta > 0$; the goal is to estimate from above the probability $\mathbb{P}\left[\gamma_t \in B_\delta(y)\right] = \mu_t[B_\delta(y)]$, so as to recover a bound on $\rho_t(y)$ as $\delta \to 0$.

If $\mathbb{P}\left[\gamma_t \in B_\delta(y)\right] = 0$, then there is nothing to prove. Otherwise we may condition γ by the event "$\gamma_t \in B_\delta(y)$". Explicitly, this means: Introduce γ' such that law $(\gamma') = \Pi' = (1_{\mathcal{Z}}\Pi)/\Pi[\mathcal{Z}]$, where

$$\mathcal{Z} = \Big\{\gamma \in \Gamma(M);\ \gamma_t \in B_\delta(y)\Big\}.$$

Further define $\pi' = \text{law}\,(\gamma'_0, \gamma'_1)$, and $\mu'_s = \text{law}\,(\gamma'_s) = (e_s)_\#\Pi'$. Obviously,

$$\Pi' \le \frac{\Pi}{\Pi[\mathcal{Z}]} = \frac{\Pi}{\mu_t[B_\delta(y)]},$$

so for all $s \in [0,1]$,

$$\mu'_s \le \frac{\mu_s}{\mu_t[B_\delta(y)]}.$$

In particular, μ'_s is absolutely continuous and its density ρ'_s satisfies (ν-almost surely)

$$\rho'_s \le \frac{\rho_s}{\mu_t[B_\delta(y)]} \tag{19.4}$$

When $s = t$, inequality (19.4) can be refined into

$$\rho'_t = \frac{\rho_t\, 1_{B_\delta(y)}}{\mu_t[B_\delta(y)]}, \tag{19.5}$$

since

$$(e_t)_\# \left(\frac{1_{\gamma_t \in B_\delta(y)}}{\mu_t[B_\delta(y)]}\right)(x) = \frac{1_{x\in B_\delta(y)}((e_t)_\#\Pi)(x)}{\mu_t[B_\delta(y)]}.$$

(This is more difficult to write down than to understand!)

From the restriction property (Theorem 4.5), (γ'_0, γ'_1) is an optimal coupling of (μ'_0, μ'_1), and therefore $(\mu'_s)_{0\le s\le 1}$ is a displacement interpolation. By Theorem 17.28 with $U(r) = -r^{1-\frac{1}{N}}$,

$$\int_M (\rho_t')^{1-\frac{1}{N}}\, d\nu \geq (1-t) \int_{M\times M} (\rho_0'(x_0))^{-\frac{1}{N}} \beta_{1-t}(x_0,x_1)^{\frac{1}{N}}\, \pi'(dx_0\, dx_1) + \; t \int_{M\times M} (\rho_1'(x_1))^{-\frac{1}{N}} \beta_t(x_0,x_1)^{\frac{1}{N}}\, \pi'(dx_0\, dx_1). \quad (19.6)$$

By definition, μ_t' is supported in $B_\delta(y)$, so

$$\int_M (\rho_t')^{1-\frac{1}{N}}\, d\nu = \int_{B_\delta(y)} (\rho_t')^{1-\frac{1}{N}}\, d\nu = \nu[B_\delta(y)] \int_{B_\delta(y)} (\rho_t')^{1-\frac{1}{N}}\, \frac{d\nu}{\nu[B_\delta(y)]}. \quad (19.7)$$

By Jensen's inequality, applied with the concave function $r \to r^{1-\frac{1}{N}}$,

$$\int_{B_\delta(y)} (\rho_t')^{1-\frac{1}{N}}\, \frac{d\nu}{\nu[B_\delta(y)]} \leq \left(\int \rho_t'\, \frac{d\nu}{\nu[B_\delta(y)]} \right)^{1-\frac{1}{N}} = \frac{1}{\nu[B_\delta(y)]^{1-\frac{1}{N}}}.$$

Plugging this in (19.7), we find

$$\int_M (\rho_t')^{1-\frac{1}{N}}\, d\nu \leq \nu[B_\delta(y)]^{\frac{1}{N}}. \quad (19.8)$$

On the other hand, from (19.4) the right-hand side of (19.6) can be bounded below by

$$\begin{aligned} &\mu_t[B_\delta(y)]^{\frac{1}{N}} \int_{M\times M} \left[(1-t)\, (\rho_0(x_0))^{-\frac{1}{N}} \beta_{1-t}(x_0,x_1)^{\frac{1}{N}} + \; t\, (\rho_1(x_1))^{-\frac{1}{N}} \beta_t(x_0,x_1)^{\frac{1}{N}} \right] \pi'(dx_0\, dx_1) \\ &= \mu_t[B_\delta(y)]^{\frac{1}{N}}\; \mathbb{E}\left[(1-t)\, (\rho_0(\gamma_0'))^{-\frac{1}{N}} \beta_{1-t}(\gamma_0',\gamma_1')^{\frac{1}{N}} + \; t\, (\rho_1(\gamma_1'))^{-\frac{1}{N}} \beta_t(\gamma_0',\gamma_1')^{\frac{1}{N}} \right] \\ &\geq \mu_t[B_\delta(y)]^{\frac{1}{N}}\; \mathbb{E} \inf_{\gamma_t \in [x_0,x_1]_t} \left[(1-t)\, (\rho_0(\gamma_0'))^{-\frac{1}{N}} \beta_{1-t}(\gamma_0',\gamma_1')^{\frac{1}{N}} + \; t\, (\rho_1(\gamma_1'))^{-\frac{1}{N}} \beta_t(\gamma_0',\gamma_1')^{\frac{1}{N}} \right], \end{aligned} \quad (19.9)$$

where the last inequality follows just from the (obvious) remark that $\gamma_t' \in [\gamma_0',\gamma_1']_t$. In all these inequalities, we can restrict π' to the set $\{\rho_0(x_0) > 0,\, \rho_1(x_1) > 0\}$ which is of full measure.

Let

$$F(x) := \inf_{x\in[x_0,x_1]_t} \left[(1-t)\, (\rho_0(x_0))^{-\frac{1}{N}} \beta_{1-t}(x_0,x_1)^{\frac{1}{N}} + \; t\, (\rho_1(x_1))^{-\frac{1}{N}} \beta_t(x_0,x_1)^{\frac{1}{N}} \right];$$

and by convention $F(x) = 0$ if either $\rho_0(x_0)$ or $\rho_1(x_1)$ vanishes. Then in view of (19.5) the lower bound in (19.9) can be rewritten as

$$\mathbb{E}\, F(\gamma_t') \; = \int_M F(x)\, d\mu_t'(x) = \frac{\displaystyle\int_{B_\delta(y)} F(x)\, d\mu_t(x)}{\mu_t[B_\delta(y)]}.$$

Now combine this with the upper bound (19.8), to conclude that

$$\left(\frac{\mu_t[B_\delta(y)]}{\nu[B_\delta(y)]} \right)^{-\frac{1}{N}} \geq \frac{\displaystyle\int_{B_\delta(y)} F(x)\, d\mu_t(x)}{\mu_t[B_\delta(y)]}. \quad (19.10)$$

Lebesgue's density theorem tells the following: if φ is a locally integrable function, then, $\nu(dy)$-almost any y is a Lebesgue point of φ, which means

$$\frac{1}{\nu[B_\delta(y)]} \int_{B_\delta(y)} \varphi(x)\, d\nu(x) \;\xrightarrow[\delta \downarrow 0]{}\; \varphi(y).$$

In particular, if y is a Lebesgue point, then

$$\frac{\mu_t[B_\delta(y)]}{\nu[B_\delta(y)]} = \frac{\displaystyle\int_{B_\delta(y)} \rho_t(x)\, d\nu(x)}{\nu[B_\delta(y)]} \;\xrightarrow[\delta \downarrow 0]{}\; \rho_t(y).$$

The inequality in (19.10) proves that $F\rho_t$ is locally ν-integrable; therefore also

$$\frac{\displaystyle\int_{B_\delta(y)} F(x)\, d\mu_t(x)}{\nu[B_\delta(y)]} \xrightarrow[\delta \downarrow 0]{} F(y)\, \rho_t(y).$$

If one plugs these two limits in (19.10), one obtains

$$\rho_t(y)^{-\frac{1}{N}} \geq \frac{F(y)\, \rho_t(y)}{\rho_t(y)} = F(y),$$

provided that $\rho_t(y) > 0$; and then $\rho_t(y) \leq F(y)^{-N}$, as desired. In the case $\rho_t(y) = 0$ the conclusion still holds true. □

It is useful to consider the particular case when the initial density μ_0 is a Dirac mass and the final mass is the uniform distribution on some set B.

Theorem 19.3 (Jacobian bounds revisited). *Let M be a Riemannian manifold equipped with a reference measure $\nu = e^{-V} \mathrm{vol}$, satisfying the curvature-dimension condition $\mathrm{CD}(K, N)$ for some $K \in \mathbb{R}$, $N \in (1, \infty)$. Let $z_0 \in M$ and let B be a bounded set of positive measure. Let further $(\mu_t^{z_0})_{0 \leq t \leq 1}$ be the displacement interpolation joining $\mu_0 = \delta_{z_0}$ to $\mu_1 = (1_B \nu)/\nu[B]$. Then the density $\rho_t^{z_0}$ of $\mu_t^{z_0}$ satisfies*

$$\rho_t^{z_0}(x) \leq \frac{C(K, N, R)}{t^N\, \nu[B]},$$

where

$$C(K, N, R) = \exp\Big(-\sqrt{(N-1)K_-}\, R\Big), \qquad K_- = \max(-K, 0), \tag{19.11}$$

and R is an upper bound on the distance between z_0 and elements of B.

In particular, if $K \geq 0$, then

$$\rho_t^{z_0}(x) \leq \frac{1}{t^N\, \nu[B]}.$$

Remark 19.4. Theorem 19.3 is a classical tool in Riemannian geometry; it is often stated as a bound on the Jacobian of the map $(s, \xi) \longmapsto \exp_x(s\xi)$. It will be a good exercise for the reader to convert Theorem 19.3 into such a Jacobian bound.

Proof of Theorem 19.3. Let z_0 and B be as in the statement of the lemma. Let $\mu_1 = (1_B\, \nu)/\nu[B]$. Consider a displacement interpolation $(\mu_t)_{0 \leq t \leq 1}$ between $\mu_0 = \delta_{z_0}$ and μ_1. Recall from Chapter 13 that μ_t is absolutely continuous for all $t \in (0, 1]$. So Theorem 19.2 can be applied to the displacement interpolation $(\mu'_t)_{0 \leq t \leq 1}$ defined by $\mu'_t = \mu_{t'}$, $t' = t_0 + (1 - t_0)t$; this yields

$$\rho_{t'}(x) \le \sup_{x\in[x_0,x_1]_t} \Big[(1-t)\,\beta_{1-t}(x_0,x_1)^{\frac{1}{N}}\,\rho_{t_0}(x_0)^{-\frac{1}{N}} + t\,\beta_t(x_0,x_1)^{\frac{1}{N}}\,\rho_1(x_1)^{-\frac{1}{N}}\Big]^{-N}. \quad (19.12)$$

Obviously, the sum above can be restricted to those pairs (x_0, x_1) such that x_1 lies in the support of μ_1, i.e. $x_1 \in B$; and x_0 lies in the support of μ_{t_0}, which implies $x_0 \in [z_0, B]_{t_0}$. Moreover, since $z \to z^{-N}$ is nonincreasing, one has the obvious bound

$$\begin{aligned}\rho_{t'}(x) &\le \sup_{x\in[x_0,x_1]_t;\ x_0\in[z_0,B]_{t_0};\ x_1\in B} \Big[t\,\beta_t(x_0,x_1)^{\frac{1}{N}}\rho_1(x_1)^{-\frac{1}{N}}\Big]^{-N} \\ &= \sup_{x\in[x_0,x_1]_t;\ x_0\in[z_0,B]_{t_0};\ x_1\in B} \frac{\rho_1(x_1)}{t^N\,\beta_t(x_0,x_1)}.\end{aligned}$$

Since $\rho_1 = 1_B/\nu[B]$, actually

$$\rho_{t'}(x) \le \frac{S(t_0,z_0,B)}{t^N \nu[B]},$$

where

$$S(t_0,z_0,B) := \sup\Big\{\beta_t(x_0,x_1)^{-\frac{1}{N}};\quad x_0\in[z_0,B]_{t_0},\ x_1\in B\Big\}. \quad (19.13)$$

Now let $t_0 \to 0$ and t go to t', in such a way that t' stays fixed. Since B is bounded, the geodesics linking z_0 to an element of B have a uniformly bounded speed, so the set $[z_0, B]_{t_0}$ is included in a ball $B(z, Vt_0)$ for some constant V; this shows that those x_0 appearing in (19.13) converge uniformly to z_0. By continuity of β_t, $S(t_0, z_0, B)$ converges to $S(0, z_0, B)$. Then an elementary estimate of β_t shows that $S(0, z_0, B) \le C(K, N, R)$. □

To conclude, I shall state a theorem which holds true with the *intrinsic* distortion coefficients of the manifold, whithout any reference to a choice of K and N, and whithout any assumption on the behavior of the manifold at infinity (if the total cost is infinite, we can appeal to the notion of generalized optimal coupling and generalized displacement interpolation, as in Chapter 13. Recall Definition 14.17.

Theorem 19.5 (intrinsic pointwise bounds on the displacement interpolant). *Let M be an n-dimensional Riemannian manifold equipped with a reference measure $\nu = e^{-V}\,\mathrm{vol}$, $V \in C^2(M)$, and let $\overline{\beta}$ be the associated distortion coefficients. Let μ_0, μ_1 be two absolutely continuous probability measures on M, let $(\mu_t)_{0\le t\le 1}$ be the unique generalized displacement interpolation between μ_0 and μ_1, and let ρ_t be the density of μ_t with respect to ν. Then one has the pointwise bound*

$$\rho_t(x) \le \sup_{x\in[x_0,x_1]_t} \left((1-t)\Big(\frac{\rho_0(x_0)}{\overline{\beta}_{1-t}(x_0,x_1)}\Big)^{-\frac{1}{n}} + t\Big(\frac{\rho_1(x_1)}{\overline{\beta}_t(x_0,x_1)}\Big)^{-\frac{1}{n}}\right)^{-n}, \quad (19.14)$$

where by convention $\big((1-t)\,a^{-\frac{1}{n}} + t\,b^{-\frac{1}{n}}\big)^{-n} = 0$ if either a or b is 0.

Proof of Theorem 19.5. First use the standard approximation procedure of Proposition 13.2 to define probability measures $\mu_{t,\ell}$ with density $\rho_{t,\ell}$, and numbers $Z\ell$ such that $Z_\ell \uparrow 1$, $Z_\ell \rho_{t,\ell} \uparrow 1$, and $\mu_{t,\ell}$ are compactly supported.

Then we can re-do the proof of Theorem 19.2 with $\mu_{0,\ell}$ and $\mu_{1,\ell}$, replacing Theorem 17.28 by Theorem 17.33. The result is

$$\rho_{t,\ell}(x) \le \sup_{x\in[x_0,x_1]_t} \left((1-t)\Big(\frac{\rho_{0,\ell}(x_0)}{\overline{\beta}_{1-t}(x_0,x_1)}\Big)^{-\frac{1}{n}} + t\Big(\frac{\rho_{1,\ell}(x_1)}{\overline{\beta}_t(x_0,x_1)}\Big)^{-\frac{1}{n}}\right)^{-n}.$$

Since $Z_\ell\, \rho_{t,\ell} \le \rho_t$, it follows that

$$Z_\ell\, \rho_{t,\ell}(x) \le \sup_{x \in [x_0,x_1]_t} \left((1-t) \Big(\frac{Z_0\, \rho_{0,\ell}(x_0)}{\overline{\beta}_{1-t}(x_0,x_1)} \Big)^{-\frac{1}{n}} + t \Big(\frac{Z_1\, \rho_{1,\ell}(x_1)}{\overline{\beta}_t(x_0,x_1)} \Big)^{-\frac{1}{n}} \right)^{-n}$$
$$\le \sup_{x \in [x_0,x_1]_t} \left((1-t) \Big(\frac{\rho_0(x_0)}{\overline{\beta}_{1-t}(x_0,x_1)} \Big)^{-\frac{1}{n}} + t \Big(\frac{\rho_1(x_1)}{\overline{\beta}_t(x_0,x_1)} \Big)^{-\frac{1}{n}} \right)^{-n}.$$

The conclusion follows by letting $\ell \to \infty$. □

Democratic condition

Poincaré inequalities are conditioned, loosely speaking, to the "richness" of the space of geodesics: one should be able to transfer mass between sets by going along geodesics, in such a way that different points use geodesics that do not get too much closer. This idea (which is reminiscent of the intuition behind the distorted Brunn–Minkowski inequality) will be more apparent in the following condition. It says that one can use geodesics to redistribute all the mass of a ball in such a way that each point in the ball sends all its mass uniformly over the ball, but no point is visited too often in the process. In the next definition, what I call "uniform distribution on B" is the reference measure ν, conditioned on the ball, that is $(1_B \nu)/\nu[B]$. The definition is formulated in the setting of a geodesic space (recall the definitions about length spaces in Chapter 7), but in this chapter we shall only apply it in smooth Riemannian manifolds.

Definition 19.6 (Democratic condition). *A measure ν on a geodesic space $(\mathcal{X}, d)$ is said to satisfy the democratic condition* $\mathrm{Dm}(C)$ *for some constant $C > 0$ if the following property holds true: For any closed ball B in $\mathcal{X}$ there is a random geodesic γ such that γ_0 and γ_1 are independent and distributed uniformly in B, and the time-integral of the density of γ_t (with respect ν) never exceeds $C/\nu[B]$.*

The condition is said to hold uniformly if the constant C is independent of the ball $B = B(x, r])$, and locally uniformly if it is independent of B as long as $B(x, 2r])$ remains inside a large fixed ball $B(z, R)$.

A more explicit formulation of the democratic condition is as follows: If μ_t stands for the law of γ_t, then

$$\int_0^1 \mu_t\, dt \le C\, \frac{\nu}{\nu[B]}. \tag{19.15}$$

Theorem 19.7 ($\mathrm{CD}(K,N)$ **implies** Dm**).** *Let M be a Riemannian manifold equipped with a reference measure $\nu = e^{-V} \mathrm{vol}$, satisfying the curvature-dimension condition* $\mathrm{CD}(K,N)$ *for some $K \in \mathbb{R}$, $N \in (1, \infty)$. Then ν satisfies a locally uniform democratic condition, with an admissible constant $2^N C(K,N,R)$ in a large ball $B(z,R)$, where $C(K,N,R)$ is defined in* (19.11).

In particular, if $K \ge 0$, then ν satisfies the uniform democratic condition $\mathrm{Dm}(2^N)$.

Proof of Theorem 19.7. The proof is largely based on Theorem 19.3.

Let B be a ball of radius r. For any point x_0, let $\mu_t^{x_0}$ be as in the statement of Theorem 19.3; then its density $\rho_t^{x_0}$ (with respect to ν) is bounded above by $C(K,N,R)/(t^N\, \nu[B])$.

On the other hand, $\mu_t^{x_0}$ can be interpreted as the position at time t of a random geodesic γ^{x_0} starting at x_0 and ending at x_1, which is distributed according to μ. By integrating

this against $\mu(dx_0)$, we obtain the position at time t of a random geodesic γ such that γ_0 and γ_1 are independent and both distributed according to μ. Explicitly,

$$\mu_t = \text{law}\,(\gamma_t) = \int_M \mu_t^{x_0}\, d\mu(x_0).$$

Obviously, the uniform bound on ρ_t persists upon integration, so

$$\mu_t \le \left[\frac{C(K,N,R)}{t^N\,\nu[B]}\right]\,\nu. \tag{19.16}$$

Recall that $\mu_t = \text{law}\,(\gamma_t)$, where γ_0, γ_1 are independent and distributed with respect to μ. Since geodesics in a Riemannian manifold are almost surely unique, we can throw away a set of zero volume in $B \times B$ such that for each $(x,y) \in (B\times B)\setminus Z$, there is a unique geodesic $(\gamma_t^{x_0,x_1})_{0\le t\le 1}$ going from x_0 to x_1. Then μ_t is characterized as the law of $\gamma_t^{x_0,x_1}$, where $\text{law}\,(x_0,x_1) = \mu\otimes\mu$. This shows that if we repeat the construction by exchanging the variables x_0 and x_1, and replacing t by $1-t$, then we get the same path (μ_t), up to reparametrization of time. So

$$\mu_t \le \left[\frac{C(K,N,R)}{(1-t)^N\,\nu[B]}\right]\,\nu. \tag{19.17}$$

Combining (19.16) and (19.17) and passing to densities, one obtains that, $\nu(dx)$-almost surely,

$$\rho_t(x) \;\le\; C(K,N,R)\min\left(\frac{1}{t^N},\,\frac{1}{(1-t)^N}\right)\,\frac{1}{\nu[B]} \;\le\; \frac{2^N\,C(K,N,R)}{\nu[B]}, \tag{19.18}$$

and Theorem 19.7 follows. □

Remark 19.8. The above bounds (19.18) can be improved as follows. Let $\mu = \rho\,\nu$ be a measure that is absolutely continuous with respect to ν, and otherwise arbitrary. Then there exists a random geodesic γ with $\text{law}\,(\gamma_0,\gamma_1) = \mu\otimes\mu$, such that $\text{law}\,(\gamma_t)$ admits a density ρ_t with respect to ν, and

$$\|\rho_t\|_{L^p} \le \min\left(\frac{1}{t^{N/p'}},\,\frac{1}{(1-t)^{N/p'}}\right)\|\rho\|_{L^p} \tag{19.19}$$

for all $p\in(1,\infty)$, where $p' = p/(p-1)$ is the conjugate exponent to p and $\|\rho\|_{L^p} = (\int \rho^p\,d\nu)^{1/p}$.

Local Poincaré inequality

Theorem 19.9 (doubling + democratic implies local Poincaré). *Let $(\mathcal{X},d)$ be a length space equipped with a reference measure ν satisfying a doubling condition with constant D, and a democratic condition with constant C. Then ν satisfies a local Poincaré inequality with constant $P = 2\,C\,D$.*

If the doubling and democratic conditions hold true inside a ball $B(z,R)$ with a constant $C = C(z,R)$ and $D = D(z,R)$ respectivley, then ν satisfies a local Poincaré inequality on the ball $B(z,R)$ with constant $P(z,R) = 2\,C(z,R)\,D(z,R)$.

Before giving the proof of Theorem 19.9 I shall state a corollary which follows immediately from this theorem together with Corollary 18.8 and Theorem 19.7:

Corollary 19.10 (Local Poincaré inequality from $\mathrm{CD}(K,N)$). *Let M be a Riemannian manifold equipped with a reference measure $\nu = e^{-V}\,\mathrm{vol}$, satisfying the curvature-dimension condition $\mathrm{CD}(K,N)$ for some $K \in \mathbb{R}$, $N \in (1,\infty)$. Then ν satisfies a local Poincaré inequality with a constant $P(K,N,R) = 2^{2N+1}\,C(K,N,R)\,D(K,N,R)$, inside any ball $B(z,R)$, where $C(K,N,R)$ and $D(K,N,R)$ are defined by (19.11) and (18.9).*

In particular, if $K \geq 0$ then ν satisfies a local Poincaré inequality on the whole of M with constant 2^{2N+1}.

Proof of Theorem 19.9. Let x_0 be a given point in M. Given $r > 0$, write $B = B_r(x_0)$, and $2B = B_{2r}(x_0)$. As before, let $\mu = (1_B\,\nu)/\nu[B]$. Let u be an arbitrary Lipschitz function. For any $y_0 \in M$, we have

$$u(y_0) \;-\; \langle u \rangle_B \;=\; \int_M (u(y_0) - u(y_1))\; d\mu(y_1). \tag{19.20}$$

Then

$$\fint_B |u - \langle u\rangle_B|\, d\nu \;=\; \int_M \big|u(y_0) - \langle u\rangle_B\big|\, d\mu(y_0) \;\leq\; \int_{B\times B} \big|u(y_0) - u(y_1)\big|\, d\mu(y_0)\, d\mu(y_1). \tag{19.21}$$

Next, we estimate $|u(y_0) - u(y_1)|$ in terms of a continuous-speed geodesic path γ joining y_0 to y_1, where $y_0, y_1 \in B$. The length of such a geodesic path is clearly less than $2r$. Then, with the shorthand $g = |\nabla u|$,

$$\big|u(y_0) - u(y_1)\big| \;\leq\; 2r \int_0^1 g(\gamma(t))\, dt. \tag{19.22}$$

By assumption there is a random geodesic γ such that $\mathrm{law}\,(\gamma_0,\gamma_1) = \mu \otimes \mu$ and $\mu_t = \mathrm{law}\,(\gamma_t)$ satisfies (19.15). Integrating (19.22) against the law of γ yields

$$\begin{aligned} \int_{M\times M} |u(y_0) - u(y_1)|\, d\mu(y_0)\, d\mu(y_1) &\leq \mathbb{E}\left(2r \int_0^1 g(\gamma(t))\, dt\right) \\ &= 2r \int_0^1 \mathbb{E}\, g(\gamma(t))\, dt \\ &= 2r \int_0^1 \int_M g\, d\mu_t\, dt. \end{aligned} \tag{19.23}$$

This, combined with (19.21), implies

$$\fint_B |u - \langle u\rangle_B|\, d\nu \leq 2r \int_0^1 \int_M g\, d\mu_t\, dt. \tag{19.24}$$

However, a geodesic joining two points in $B(x,r)$ cannot leave $B(x,2r)$, so (19.24) and the democratic condition together imply that

$$\fint_B |u - \langle u\rangle_B|\, d\nu \leq \frac{2\,C\,r}{\nu[B]} \int_{2B} g\, d\nu. \tag{19.25}$$

By the doubling property, $\frac{1}{\nu[B]} \;\leq\; \frac{D}{\nu[2B]}$. The conclusion is that

$$\fint_B |u - \langle u\rangle_B|\, d\nu \;\leq\; 2\,C\,D\,r \fint_{2B} g\, d\nu. \tag{19.26}$$

This concludes the proof of Theorem 19.9. □

Remark 19.11. With almost the same proof, it is easy to derive the following refinement of the local Poincaré inequality:

$$\int_{B(x,r)} \frac{|u(x)-u(y)|}{d(x,y)}\, d\nu(x)\, d\nu(y) \;\le\; P(K,N,R) \int_{B(x,2r)} |\nabla u|(x)\, d\nu(x).$$

Back to Brunn–Minkowski and Prékopa–Leindler inequalities

To conclude this chapter I shall explain how the Brunn–Minkowski inequality (18.4) follows at once from the pointwise estimates on the interpolant density.

Proof of Theorem 18.4, again. Let μ_0 be the measure ν conditioned on A_0, i.e. $\mu_0 = \rho_0\, \nu$ with $\rho_0 = 1_{A_0}/\nu[A_0]$. Similarly, let $\mu_1 = \rho_1\, \nu$ with $\rho_1 = 1_{A_1}/\nu[A_1]$. Let ρ_t be the density of the displacement interpolant at time t. Then, since ρ_0 vanishes out of A_0, and ρ_1 out of A_1, Theorem 19.2 yields

$$\begin{aligned}\rho_t(x)^{-\frac{1}{N}} &\ge (1-t) \left[\inf_{x\in[A_0,A_1]_t} \beta_{1-t}(x_0,x_1)^{\frac{1}{N}}\right] \nu[A_0]^{\frac{1}{N}} \;+\; \left[\inf_{x\in[A_0,A_1]_t} \beta_t(x_0,x_1)^{\frac{1}{N}}\right] \nu[A_1]^{\frac{1}{N}} \\ &\ge (1-t) \left[\inf_{(x_0,x_1)\in A_0\times A_1} \beta_{1-t}(x_0,x_1)^{\frac{1}{N}}\right] \nu[A_0]^{\frac{1}{N}} \;+\; \left[\inf_{(x_0,x_1)\in A_0\times A_1} \beta_t(x_0,x_1)^{\frac{1}{N}}\right] \nu[A_1]^{\frac{1}{N}}.\end{aligned}$$

Now integrate this against $\rho_t(x)\, d\nu(x)$: since the right-hand side does not depend on x any longer, it follows that

$$\begin{aligned}\int \rho_t(x)^{1-\frac{1}{N}}\, d\nu(x) \ge (1-t) &\left[\inf_{x\in[A_0,A_1]_t} \beta_{1-t}(x_0,x_1)^{\frac{1}{N}}\right] \nu[A_0]^{\frac{1}{N}} \\ &+ \left[\inf_{x\in[A_0,A_1]_t} \beta_t(x_0,x_1)^{\frac{1}{N}}\right] \nu[A_1]^{\frac{1}{N}}.\end{aligned}$$

On the other hand, ρ_t is concentrated on $[A_0,A_1]_t$, so the same Jensen inequality that was used in the earlier proof of Theorem 18.4 implies

$$\int \rho_t(x)^{1-\frac{1}{N}}\, d\nu(x) \le \nu\big[[A_0,A_1]_t\big]^{\frac{1}{N}},$$

and inequality (18.4) follows. □

Now it is interesting to see that Theorem 19.2 also implies the **distorted Prékopa–Leindler inequality**. Theorem 19.2. This is a functional variant of the Brunn–Minkowski inequality, which is sometimes much more convenient to handle. (Here I say that the inequality is "distorted" only because the Prékopa–Leindler inequality is usually stated in $\mathbb{R}^n$, while the Riemannian generalization involves distortion coefficients.) I shall first consider the dimension-free case, which is simpler and does not need distortion coefficients.

Theorem 19.12 (Prékopa–Leindler inequalities). *With the same notation as in Theorem 19.2, assume that (M,ν) satisfies the curvature-dimension condition $\mathrm{CD}(K,\infty)$. Let $t\in(0,1)$, and let f, g and h be three nonnegative functions with finite moments of order p, such that the inequality*

$$h(x) \ge \sup_{x\in[x_0,x_1]_t} f(x_0)^{1-t}\, g(x_1)^t\, \exp\left(-\frac{K\,t(1-t)}{2}\, d(x_0,x_1)^2\right) \tag{19.27}$$

is satisfied for all $x\in M$. Then

$$\int h\, d\nu \ge \left(\int f\, d\nu\right)^{1-t} \left(\int g\, d\nu\right)^t.$$

Proof of Theorem 19.12. By homogeneity, one may assume $\int f = \int g = 1$. Write then $\rho_0 = f$, $\rho_1 = g$; by Theorem 19.2, the displacement interpolant ρ_t between $\rho_0\,\nu$ and $\rho_1\,\nu$ satisfies (19.3). From (19.27), $h \geq \rho_t$. It follows that $\int h \geq \int \rho_t = 1$, as desired. □

The following corollary will be useful later:

Corollary 19.13 (Positive curvature implies square-exponential moments). *Let (M, ν) be a Riemannian manifold satisfying a curvature-dimension bound $\mathrm{CD}(K, \infty)$ with $K > 0$. Then for any $x_0 \in X$ there exists $\lambda > 0$ such that*

$$\int e^{\lambda\, d(x_0,x)^2}\, d\nu(x) < \infty.$$

Remark 19.14. In Chapter 30 we shall see another proof of this result, showing that any $\lambda < K$ is admissible (see Theorem 30.13).

Proof of Corollary 19.13. Let A be compact such that $\nu[A] > 0$. Apply the Prékopa–Leindler inequality with $t = 1/2$, $f = 1_A$, $g = \exp(K\, d(x, A)^2/4)$ and $h = 1$. This shows that

$$\int_M e^{\frac{K\, d(x,A)^2}{4}}\, d\nu(x) < \infty, \tag{19.28}$$

and the desired conclusion follows easily. □

I shall conclude with the dimension-dependent form of the Prékopa–Leindler inequality, which will require some more notation. For any $a, b \geq 0$, $t \in [0, 1]$, $q \in \mathbb{R} \setminus \{0\}$, define

$$\mathcal{M}_t^q(a, b) := \Big[(1 - t)\, a^q \;+\; t\, b^q\Big]^{\frac{1}{q}},$$

with the convention that $\mathcal{M}_t^q(a, b) = 0$ if either a or b is 0; and $\mathcal{M}_t^{-\infty}(a, b) = \min(a, b)$.

Theorem 19.15 (dimension-dependent distorted Prékopa–Leindler inequality). *With the same notation and assumptions as in Theorem 19.2, assume that ν satisfies a curvature-dimension bound $\mathrm{CD}(K, N)$, for some $K \in \mathbb{R}$, $N \in (1, \infty)$. Let f, g and h be three nonnegative functions on M satisfying*

$$h(x) \geq \sup_{x \in [x_0,x_1]_t} \mathcal{M}_t^q\left(\frac{f(x_0)}{\beta_{1-t}^{(K,N)}(x_0, x_1)}, \frac{g(x_1)}{\beta_t^{(K,N)}(x_0, x_1)}\right), \qquad q \geq -\frac{1}{N}; \tag{19.29}$$

then

$$\int h\, d\nu \geq \mathcal{M}_t^{\frac{q}{1+Nq}}\left(\int f\, d\nu,\ \int g\, d\nu\right). \tag{19.30}$$

Proof of Theorem 19.15. The proof is quite similar to the proof of Theorem 19.12, except that now N is finite. Let f, g and h satisfy the assumptions of the theorem, define $\rho_0 = f/\|f\|_{L^1}$, $\rho_1 = g/\|g\|_{L^1}$, and let ρ_t be the density of the displacement interpolant at time t between $\rho_0\,\nu$ and $\rho_1\,\nu$. Let $\mathcal{M}$ be the right-hand side of (19.30); the problem is to show that $\int (h/\mathcal{M}) \geq 1$, and this is obviously true if $h/\mathcal{M} \geq \rho_t$. In view of Theorem 19.2, it is sufficient to establish

$$\frac{h(x)}{\mathcal{M}} \geq \sup_{x_0,x_1} \mathcal{M}_t^{\frac{1}{N}}\left(\frac{\beta_{1-t}(x_0, x_1)}{\rho_0(x_0)},\ \frac{\beta_t(x_0, x_1)}{\rho_1(x_1)}\right)^{-1}. \tag{19.31}$$

In view on the assumption of h and the form of M, it is sufficient to check that

$$\frac{1}{\mathcal{M}_t^{\frac{1}{N}}\left(\frac{\beta_{1-t}(x_0,x_1)}{\rho_0(x_0)},\ \frac{\beta_t(x_0,x_1)}{\rho_1(x_1)}\right)} \le \frac{\mathcal{M}_t^q\left(\frac{f(x_0)}{\beta_{1-t}(x_0,x_1)},\ \frac{g(x_1)}{\beta_t(x_0,x_1)}\right)}{\mathcal{M}_t^{\frac{q}{1+Nq}}(\|f\|_{L^1},\ \|g\|_{L^1})}.$$

But this is a consequence of the following computation:

$$\frac{1}{\mathcal{M}_t^{-s}(a^{-1},b^{-1})} = \mathcal{M}_t^s(a,b) \le \mathcal{M}_t^q\left(\frac{a}{c},\ \frac{b}{d}\right)\mathcal{M}_t{}^r(c,d) = \frac{\mathcal{M}_t^q\left(\frac{a}{c},\frac{b}{d}\right)}{\mathcal{M}_t^{-r}(c,d)}, \tag{19.32}$$

$$\frac{1}{q}+\frac{1}{r}=\frac{1}{s}, \qquad q+r\ge 0,$$

where the two equalities in (19.32) are obvious by homogeneity, and the central inequality is a consequence of the two-point Hölder inequality (see the bibliographical notes for references). □

Bibliographical Notes

The main historical references about interior regularity estimates are by De Giorgi [129], Nash [281] and Moser [277, 278]. Their methods were later considerably developed in the theory of elliptic partial differential equations, see e.g. [185, 92]. Moser's Harnack inequality is a handy technical tool to recover the most famous regularity results. The relations of this inequality with Poincaré and Sobolev inequalities, and the influence of Ricci curvature on it, were studied by many authors including in particular Saloff-Coste [326] and Grigor'yan [192].

Lebesgue's density theorem can be found in most textbooks about measure theory, e.g. Rudin [315, Chapter 7].

Local Poincaré inequalities admit many variants and are known under many names in the literature, in particular "weak Poincaré inequalities", by contrast with "strong" Poincaré inequalities, in which the larger ball is not $B(x,2r)$ but $B(x,r)$. In spite of that terminology, both inequalities are in some sense equivalent [206]. Sometimes one replaces the ball $B(x,2r)$ by a smaller ball $B(x,\lambda r)$, $\lambda>1$. One also says sometimes that the inequality (19.1) is of type $(1,1)$ because there are L^1 norms on both sides. Inequality (19.1) also implies the other main members of the family of local Poincaré inequalities, see for instance Heinonen [208, Chapters 4 and 9]. There are equivalent formulations of these inequalities in terms of modulus and capacity, see e.g. [224, 225] and the many references therein. The study of Poincaré inequalities in metric spaces has turned into a surprisingly large domain of research.

Theorem 19.3 is a classical estimate, usually formulated in terms of Jacobian estimates, see e.g. Saloff-Coste [327, p. 179]; there are some differences in the formulas due to the convention that geodesics might be parametrized by arc length rather than defined on $[0,1]$. A transport-based proof was devised by Lott and myself in [249].

The "intrinsic" bounds appearing in Theorem 19.5 go back to [118] (in the compactly supported case) and [**?**] (in the general case). The methods used in these references are different. The restriction strategy which I used for Theorems 19.2 and 19.5 is an amplification of the transport-based proof of Theorem 19.3 appearing in [249].

The democratic condition $\mathrm{Dm}(C)$ was explicitly introduced in [249], but it is somehow implicit in previous works, such as Cheeger and Colding [111]. The proofs of Theorems 19.7 and 19.9 closely follow [249]. It was Pajot who pointed out to me the usefulness of Jacobian

estimates expressed by Theorem 19.3 (recall Remark 19.4) for proving local Poincaré inequalities.

Proofs of the local Poincaré inequality from optimal transport, based on slightly different but quite close arguments, were found independently by von Renesse [368] and Sturm [339].

The classical Prékopa–Leindler inequality in Euclidean space goes back to [303, 238]; see [181] for references and its role in the Brunn–Minkowski theory. Although in principle equivalent to the Brunn–Minkowski inequality, it is sometimes quite more handy, see e.g. a famous application by Maurey [262] to concentration inequalities. Bobkov and Ledoux [59] have shown how to use this inequality to derive many functional inequalities such as logarithmic Sobolev inequalities, to be considered in Chapter 21.

In the Euclidean case, the stronger version of the Prékopa–Leindler inequality which corresponds to Theorem 19.15 was established by Borell [65], Brascamp and Lieb [68], and others. The proof of Theorem 19.15 from Theorem 19.2 follows the argument given at the very end of [118]. The inequality used in (19.32) appears in [181, Lemma 10.1].

The Prékopa–Leindler inequality on manifolds, Theorem 19.12, appears in recent work by Cordero-Erausquin, McCann and Schmuckenschläger [117]. In that paper displacement convexity is established independently of the Prékopa–Leindler inequality, but with similar tools (namely, the Jacobian estimates in Chapter 14). The presentation that I have followed makes it clear that the Prékopa–Leindler inequality, and even the stronger pointwise bounds in Theorem 19.2, can really be seen as a consequence of displacement convexity inequalities (together with the restriction property). This obstination to derive everything from displacement convexity, rather than directly from Jacobian estimates, will find a justification in Part III of these notes: In some sense displacement convexity is a quite softer and more robust notion, than Jacobian estimates.

In $\mathbb{R}^N$, there is also a "stronger" version in which the exponent q can go down to $-1/(N-1)$ instead of $-1/N$; it reads

$$h\big((1-t)x_0+tx_1\big) \geq \mathcal{M}_t^q(f(x_0),g(x_1)) \;\Longrightarrow\; \int h(z)\,dz \geq \mathcal{M}_t^{\frac{q}{1+q(N-1)}}(m_i(f),m_i(g))\cdot\mathcal{M}_t^1\left(\frac{1}{m_i(f)}\int f,\ \frac{1}{m_i(g)}\right. \tag{19.33}$$

It was recently shown by Bobkov and Ledoux [**?**] that this inequality can be used to establish optimal Sobolev inequalities in $\mathbb{R}^N$ (with the usual Prékopa–Leindler inequality one can apparently reach only the logarithmic Sobolev inequality, that is, the dimension-free case). See [**?**] for the history and derivation of (19.33).

20

Infinitesimal displacement convexity

The goal of the present chapter is to translate convexity inequalities of the form "the graph of a convex function lies below the chord" into inequalities of the form "the graph of a convex function lies above the tangent" — just as in statements (ii) and (iii) of Proposition 16.2. This corresponds to the limit $t \to 0$ in the convexity inequality.

One of the main results in this chapter is the HWI inequality, Theorem 20.7 below.

Time-derivative of the energy

As a preliminary step, a useful lower bound will now be given for a derivative of $U_\nu(\mu_t)$, where $(\mu_t)_{0 \le t \le 1}$ is a Wasserstein geodesic and U_ν an energy functional with a reference measure ν. This computation hardly needs any regularity on the space, so I will present it on an arbitrary length space.

In the next theorem, I consider a locally compact, complete length space $\mathcal{X}$ equipped with a distance d and a locally finite measure ν. Then $U : [0, +\infty) \to \mathbb{R}$ is a continuous convex function, twice differentiable on $(0, +\infty)$; to U is associated the functional

$$U_\nu(\mu) = \int_{\mathcal{X}} U(\rho)\, d\nu \qquad \mu = \rho\, \nu,$$

well-defined on $P_2^{\rm ac}(\mathcal{X})$.

The statement below will involve norms of gradients. In a nonsmooth length space, there is no natural notion for the gradient ∇f of a function f, but there are natural notions for the *norm* of the gradient, $|\nabla f|$. The most common one is

$$|\nabla f|(x) := \limsup_{y \to x} \frac{[f(y) - f(x)]}{d(x,y)}. \tag{20.1}$$

Rigorously speaking, this formula makes sense only if x is not isolated, which will always be the case in the sequel. A slightly finer notion is the following:

$$|\nabla^- f|(x) := \limsup_{y \to x} \frac{[f(y) - f(x)]_-}{d(x,y)}, \tag{20.2}$$

where $a_- = \max(-a, 0)$ stands for the negative part of a (which is a nonnegative number!!). It is obvious that $|\nabla^- f| \le |\nabla f|$, and both notions coincide with the usual one if f is differentiable. Note that $|\nabla^- f|(x)$ is automatically 0 if x is a local minimum of f.

Theorem 20.1 (differentiating an energy along optimal transport). *Let $(\mathcal{X}, d, \nu)$ and U be as above, and let $(\mu_t)_{0\le t\le 1}$ be a geodesic in $P_2(\mathcal{X})$, such that each μ_t is absolutely continuous with respect to ν, with density ρ_t, and $U(\rho_t)_-$ is ν-integrable for all t. Further assume that ρ_0 is Lipschitz continuous, $U(\rho_0)$ and $\rho_0\, U'(\rho_0)$ are ν-integrable, and U' is Lipschitz continuous on $\rho_0(\mathcal{X})$. Then*

$$\liminf_{t\downarrow 0} \left[\frac{U_\nu(\mu_t) - U_\nu(\mu_0)}{t}\right] \ge -\int_{\mathcal{X}} U''(\rho_0(x_0))|\nabla^- \rho_0|(x_0)\, d(x_0,x_1)\, \pi(dx_0\, dx_1), \qquad (20.3)$$

where π is an optimal coupling of (μ_0,μ_1) associated with the geodesic path $(\mu_t)_{0\le t\le 1}$.

Remark 20.2. The technical assumption on the negative part of $U(\rho_t)$ being integrable is a standard way to make sure that $U_\nu(\mu_t)$ is well-defined, with values in $\mathbb{R}\cup\{+\infty\}$. As for the assumption about U' being Lipschitz on $\rho_0(\mathcal{X})$, it means in practice that either U is twice (right-)differentiable at the origin, or ρ_0 is bounded away from 0.

Remark 20.3. Here is a more probabilistic reformulation of (20.3) (which will also make more explicit the link between π and μ_t): Let γ be a random geodesic such that $\mu_t = \text{law}\,(\gamma_t)$, then

$$\liminf_{t\downarrow 0} \left[\frac{U_\nu(\mu_t) - U_\nu(\mu_0)}{t}\right] \ge -\mathbb{E}\left[U''(\rho_0(\gamma_0))|\nabla^-\rho_0|(\gamma_0)\, d(\gamma_0,\gamma_1)\right].$$

Proof of Theorem 20.1. By convexity,

$$U(\rho_t) - U(\rho_0) \ge U'(\rho_0)\ (\rho_t - \rho_0), \qquad (20.4)$$

where $U'(0)$ is the right-derivative of U at 0.

On one hand, $U(\rho_0)$ and $U(\rho_t)_-$ are ν-integrable by assumption, so that the integral of the left-hand side of (20.4) makes sense in $\mathbb{R}\cup\{+\infty\}$ (and the integral of each term is well-defined). On the other hand, $\rho_0\, U'(\rho_0)$ is integrable by assumption, while $\rho_t\, U'(\rho_0)$ is bounded above by $(\max U')\rho_t$, which is integrable; so the integral of the right-hand side makes sense in $\mathbb{R}\cup\{-\infty\}$. All in all, inequality (20.4) can be integrated into

$$\begin{aligned} U_\nu(\mu_t) - U_\nu(\mu_0) &\ge \int U'(\rho_0)\rho_t\, d\nu - \int U'(\rho_0)\rho_0\, d\nu \\ &= \int U'(\rho_0)\, d\mu_t - \int U'(\rho_0)\, d\mu_0. \end{aligned}$$

Now let γ be a random geodesic, such that $\mu_t = \text{law}\,(\gamma_t)$. Then the above inequality can be rewritten

$$U_\nu(\mu_t) - U_\nu(\mu_0) \ge \mathbb{E}\, U'(\rho_0(\gamma_t)) \ - \ \mathbb{E}\, U'(\rho_0(\gamma_0)) = \mathbb{E}\left[U'(\rho_0(\gamma_t)) - U'(\rho_0(\gamma_0))\right].$$

Since U' is nondecreasing,

$$U'(\rho_0(\gamma_t)) - U'(\rho_0(\gamma_0)) \ge \left[U'(\rho_0(\gamma_t)) - U'(\rho_0(\gamma_0))\right]\, 1_{\rho_0(\gamma_0)>\rho_0(\gamma_t)}.$$

Multiplying and dividing by $\rho_0(\gamma_t) - \rho_0(\gamma_0)$, and then by $d(\gamma_0,\gamma_t)$, one arrives at

$$\left[U_\nu(\mu_t) - U_\nu(\mu_0)\right] \ge \mathbb{E}\left(\frac{U'(\rho_0(\gamma_t)) - U'(\rho_0(\gamma_0))}{\rho_0(\gamma_t) - \rho_0(\gamma_0)}\right)\left(\frac{\rho_0(\gamma_t) - \rho_0(\gamma_0)}{d(\gamma_0,\gamma_t)}\, 1_{\rho_0(\gamma_0)>\rho_0(\gamma_t)}\right) d(\gamma_0,\gamma_t).$$

After division by t and use of the identity $d(\gamma_0,\gamma_t) = t\, d(\gamma_0,\gamma_1)$, one obtains in the end

$$\frac{1}{t}\Big[U_\nu(\mu_t) - U_\nu(\mu_0)\Big] \geq$$
$$\mathbb{E}\left(\frac{U'(\rho_0(\gamma_t)) - U'(\rho_0(\gamma_0))}{\rho_0(\gamma_t) - \rho_0(\gamma_0)}\right)\left(\frac{\rho_0(\gamma_t) - \rho_0(\gamma_0)}{d(\gamma_0,\gamma_t)}\,1_{\rho_0(\gamma_0)>\rho_0(\gamma_t)}\right) d(\gamma_0,\gamma_1). \quad (20.5)$$

It remains to pass to the limit in the right-hand side of (20.5) as $t \to 0$. Since ρ_0 is continuous, for each geodesic γ one has $\rho_0(\gamma_t) \to \rho_0(\gamma_0) > 0$ as $t \to 0$, and in particular,

$$\frac{U'(\rho_0(\gamma_t)) - U'(\rho_0(\gamma_0))}{\rho_0(\gamma_t) - \rho_0(\gamma_0)} \xrightarrow[t\to 0]{} U''(\rho_0(\gamma_0)),$$

Similarly,

$$\liminf_{t\to 0}\left(\frac{\rho_0(\gamma_t) - \rho_0(\gamma_0)}{d(\gamma_0,\gamma_t)}\,1_{\rho_0(\gamma_0)>\rho_0(\gamma_t)}\right) \geq -|\nabla^- \rho_0|(\gamma_0).$$

So, if $v_t(\gamma)$ stands for the integrand in the right-hand side of (20.5), one has

$$\liminf v_t(\gamma) \geq -\,U''(\rho_0(\gamma_0))\,|\nabla^-\rho_0|(\gamma_0)\,d(\gamma_0,\gamma_1).$$

On the other hand, ρ_0 is Lipschitz by assumption, and also U' is Lipschitz on the range of ρ_0. So $|v_t(\gamma)| \leq C d(\gamma_0,\gamma_1)$, where C is the product of the Lipschitz constants of ρ_0 and U'. This uniform domination makes it possible to apply Fatou's lemma, in the form $\liminf_{t\to 0} \mathbb{E}\, v_t(\gamma) \geq \mathbb{E}\,\liminf v_t(\gamma)$. This translates into

$$\liminf_{t\to 0} \frac{1}{t}\Big[U_\nu(\mu_t) - U_\nu(\mu_0)\Big] \geq -\mathbb{E}\,U''(\rho_0(\gamma_0))\,|\nabla^-\rho_0|(\gamma_0)\,d(\gamma_0,\gamma_1),$$

as desired. □

Remark 20.4. This theorem does not assume smoothness of $\mathcal{X}$, and does not either assume structural restrictions on the function U. On the other hand, when $\mathcal{X}$ is a Riemannian manifold of dimension n with adequate assumptions at infinity (asymptoticall flat, in the sense of Definition **??**), and U lies in $\mathcal{DC}_n$, then there is a more precise result:

$$\lim_{t\to 0} \frac{[U_\nu(\mu_t) - U_\nu(\mu_0)]}{t} = -\int p(\rho_0)\Delta\psi, \quad (20.6)$$

where ψ is such that $T = \exp(\nabla\psi)$ is the unique optimal transport from μ_0 to μ_1, and $\Delta\psi$ is the Laplacian of ψ, defined almost everywhere; as usual, the reference measure is the standard volume measure. It is not clear a priori how this compares with the result of Theorem 20.1, but then, under slightly more stringent regularity assumptions, one can justify the integration by parts formula

$$-\int p(\rho_0)\,\Delta\psi \geq \int \rho_0\,U''(\rho_0)\nabla\rho_0\cdot\nabla\psi \quad (20.7)$$

(note indeed that $p'(r) = rU''(r)$). Since π is of the form $(\rho_0\,\nu)\otimes\delta_{x_1=T(x_0)}$ with $T = \exp\nabla\psi$, the right-hand side can be rewritten

$$\int U''(\rho_0)\nabla\rho_0\cdot\nabla\psi\,d\pi.$$

Since $|\nabla\psi(x_0)| = d(x_0,x_1)$, this integral is obviously an upper bound for the expression in (20.3). See the bibliographical notes for further comments. In the present chapter, the more precise result (20.6) will not be useful, but later in Chapter 23 we shall have to go through it.

Exercise 20.5. Use Otto's calculus to guess that $(d/dt)U_\nu(\mu_t)$ should coincide with the right-hand side of (20.7).

Dimension-free inequalities

Recall from Chapters 16 and 17 that the dimension-independent bound $\mathrm{CD}(K,\infty)$ results in convexity properties for certain functionals. A particular case of Theorem 17.15 is as follows: $\mathrm{CD}(K,\infty)$ implies the λ_U-displacement convexity of U_ν, for all $U \in \mathcal{DC}_\infty$, where

$$\lambda_U = \inf_{r\in\mathbb{R}_+} \frac{Kp(r)}{r}, \qquad p(r) = rU'(r) - U(r). \tag{20.8}$$

In particular, the Boltzmann functional H_ν is K-displacement convex.

Let $(\mu_t)_{0\le t\le 1}$ be a geodesic in $P_2^{\mathrm{ac}}(M)$, for $t > 0$, rewrite the convexity inequality

$$U_\nu(\mu_t) \le (1-t)\, U_\nu(\mu_0) + t\, U_\nu(\mu_1) - \frac{\lambda_U\, t(1-t)}{2}\, W_2(\mu_0,\mu_1)^2$$

as

$$\frac{U_\nu(\mu_t) - U_\nu(\mu_0)}{t} \le U_\nu(\mu_1) - U_\nu(\mu_0) - \frac{\lambda_U\,(1-t)}{2}\, W_2(\mu_0,\mu_1)^2.$$

Under suitable assumptions, one can apply Theorem 20.1 to pass to the limit as $t \to 0$, and obtain

$$-\int_M U''(\rho_0)\, |\nabla^- \rho_0|\, d(x_0,x_1)\, \pi(dx_0\, dx_1) \le U_\nu(\mu_1) - U_\nu(\mu_0) - \frac{\lambda_U\, W_2(\mu_0,\mu_1)^2}{2}.$$

To summarize the above computations, we have the following statement:

Theorem 20.6 (pre-HWI inequalities). *Let M be a Riemannian manifold equipped with a reference measure $\nu = e^{-V}\mathrm{vol}$, such that $\mathrm{Ric}_{\infty,\nu} \ge K$. Let $U \in \mathcal{DC}_\infty$, and let λ_U be defined by (20.8). Then, for any Wasserstein geodesic $(\mu_t = \rho_t\,\nu)_{0\le t\le 1}$ satisfying the assumptions of Theorem 20.1, one has the inequality*

$$U_\nu(\mu_0) + \frac{\lambda_U\, W_2(\mu_0,\mu_1)^2}{2} \le U_\nu(\mu_1) + \int U''(\rho_0(x_0))\, |\nabla^-\rho_0|(x_0)\, d(x_0,x_1)\, \pi(dx_0\, dx_1).$$

Many useful functional inequalities can be recovered from Theorem 20.6. Here below is the most characteristic example, called "HWI inequality"; it is expressed in terms of

- the H-functional of Boltzmann, $H_\nu(\mu) = \int \rho \log\rho\, d\nu$ (as usual $\rho = d\mu/d\nu$);
- the Wasserstein distance of order 2, W_2,
- the Fisher information I, defined by $I_\nu(\mu) = \int \frac{|\nabla\rho|^2}{\rho}\, d\nu$.

Theorem 20.7 (HWI inequality). *Let M be a Riemannian manifold equipped with a reference probability measure $\nu = e^{-V}\mathrm{vol}$, satisfying the curvature condition $\mathrm{Ric}_{\infty,\nu} \ge K$ ($K \in \mathbb{R}$). Let further μ_0 and μ_1 be two probability measures in $P_2(M)$. Then*

$$H_\nu(\mu_0) \le H_\nu(\mu_1) + W_2(\mu_0,\mu_1)\sqrt{I_\nu(\mu_0)} - \frac{K\, W_2(\mu_0,\mu_1)^2}{2}. \tag{20.9}$$

In particular, if $\nu \in P_2(M)$, then, for all $\mu \in P_2(M)$,

$$H_\nu(\mu) \le W_2(\mu,\nu)\sqrt{I_\nu(\mu)} - \frac{K\, W_2(\mu,\nu)^2}{2}. \tag{20.10}$$

Remark 20.8. The functional H_ν can be extended into a lower semi-continuous convex functional on the whole of $P_2(M)$ (just set $H_\nu(\mu) = +\infty$ if μ is not absolutely continuous with respect to ν). It is also possible to extend I_ν into a lower semi-continuous functional on the whole of $P_2(M)$; so the statement of Theorem 20.7 makes sense as it is. However, I shall prove it only in the case when μ_0 and μ_1 are absolutely continuous, and the density ρ_0 of μ_0 is Lipschitz and bounded away from 0 and ∞ (i.e. there exists $\varepsilon > 0$ such that $\varepsilon \le \rho_0 \le \varepsilon^{-1}$). Then the general case follows by a more or less standard approximation argument (see the bibliographic notes for references).

Remark 20.9. The HWI inequality plays the role of a nonlinear interpolation inequality: it shows that the Kullback information H is controlled by a bit of the Fisher information I (which is stronger, in the sense that it involves smoothness) and the Wasserstein distance W_2 (which is weaker). A related "linear" inequality is $\|h\|_{L^2} \le \sqrt{\|h\|_{H^{-1}}\, \|h\|_{H^1}}$, where H^1 is the Sobolev space defined by the L^2-norm of the gradient, and H^{-1} is the dual of H^1.

Proof of Theorem 20.7. First of all, it is clear that (20.10) follows from (20.9) upon choosing $\mu = \mu_0$, $\mu_1 = \nu$.

As explained in Remark 20.8, I shall assume that $\mu_0 = \rho_0 \nu$, $\mu_1 = \rho_1 \nu$, with ρ_0 Lipschitz and bounded away from 0 and ∞. Then $(\rho_0 \log \rho_0)_-$ and $(\rho_1 \log \rho_1)_-$ are integrable, according to Theorem 17.8. If $(\rho_1 \log \rho_1)_+$ is not integrable, then there is nothing to prove, so we might assume $H_\nu(\mu_1) < +\infty$. Then $(\rho_0 \log \rho_0)_+$ is bounded by $[\sup(\log \rho_0)_+]\rho_0$, which is integrable. So $\rho_0 \log \rho_0$ is integrable. Moreover, since ρ_0 is bounded away from 0 and ∞, the function $r \longmapsto \log r + 1$ is bounded and Lipschitz on $\rho_0(M)$. So ρ_0 satisfies all the assumptions required in Theorem 20.1. Let then $(\mu_t)_{0 \le t \le 1}$ be the unique Wasserstein geodesic joining μ_0 to μ_1; we know from Theorem 8.5(ii) that each μ_t is absolutely continuous. So all the assumptions of Theorem 20.1 are satisfied, and as a consequence one can apply Proposition 20.6 to get

$$H_\nu(\mu_0) + \frac{K\, W_2(\mu_0, \mu_1)^2}{2} \;\le\; H_\nu(\mu_1) + \int_M \frac{|\nabla^- \rho_0|(x_0)}{\rho_0(x_0)}\, d(x_0, x_1)\, \pi(dx_0\, dx_1).$$

Then, by Cauchy–Schwarz inequality,

$$\int_M \frac{|\nabla^- \rho_0|(x_0)}{\rho_0(x_0)}\, d(x_0, x_1)\, \pi(dx_0\, dx_1) \;\le\; \sqrt{\int_M d(x_0, x_1)^2\, \pi(dx_0\, dx_1)}\, \sqrt{\int_M \frac{|\nabla^- \rho_0|^2(x_0)}{\rho_0(x_0)^2}\, \pi(dx_0\, dx_1)}.$$

But by the marginal property,

$$\int_M \frac{|\nabla^- \rho_0|^2(x_0)}{\rho_0(x_0)^2}\, \pi(dx_0\, dx_1) = \int_M \frac{|\nabla^- \rho_0|^2(x_0)}{\rho_0(x_0)^2}\, \mu(dx_0) = \int_M \frac{|\nabla^- \rho_0|^2(x_0)}{\rho_0(x_0)}\, \nu(dx_0).$$

This concludes the proof of (20.9). □

I shall conclude this chapter with a generalization of the HWI inequality for functions in $\mathcal{DC}_N$.

Theorem 20.10 (Generalized HWI inequality). *Let M be a Riemannian manifold, equipped with a reference probability measure $\nu = e^{-V} \mathrm{vol}$, satisfying the curvature condition $\mathrm{CD}(K, N)$ ($K \ge 0$), and let $U \in \mathcal{DC}_N$. Let further μ_0 and μ_1 be two probability measures in $P_2(M)$. Then*

$$U_\nu(\mu_0) \le U_\nu(\mu_1) + W_2(\mu_0,\mu_1)\sqrt{I_\nu(\mu_0)} - \frac{K\|\rho_0,\rho_1\|_{L^\infty}^{\frac{1}{N}}\, W_2(\mu_0,\mu_1)^2}{2}, \tag{20.11}$$

where

$$\|\rho_0,\rho_1\|_{L^\infty} = \max\Big(\|\rho_0\|_{L^\infty(\nu)},\, \|\rho_1\|_{L^\infty(\nu)}\Big).$$

Proof of Theorem 20.10. This is the same proof as before, taking into account the result of Exercise 17.20. □

Dimension-dependent inequalities

For the kind of inequalities that are considered in this and the following chapters, it is notoriously difficult to take dimension into account properly. As a matter of fact, so far nobody has found a clean analogue of Theorem 20.7 if the assumption $\mathrm{CD}(K,\infty)$ is replaced by $\mathrm{CD}(K,N)$. Here I shall discuss a partial result in this direction.

Let me start by recalling some notation about the "distorted displacement convexity" considered at the end of Chapter 17. Given $K \in \mathbb{R}$ and $N > 1$, define the distortion coefficient $\beta_t(x_0,x_1) = \beta_t^{(K,N)}(x_0,x_1)$ as in (14.60):

$$\beta_t^{(K,N)}(x_0,x_1) \;=\; \begin{cases} \infty & \text{if } K>0 \text{ and } \alpha > \pi, \\ \left(\frac{\sin(t\alpha)}{t\sin\alpha}\right)^{N-1} & \text{if } K>0 \text{ and } \alpha \in [0,\pi], \\ 1 & \text{if } K = 0, \\ \left(\frac{\sinh(t\alpha)}{t\sinh\alpha}\right)^{N-1} & \text{if } K<0, \end{cases} \tag{20.12}$$

where

$$\alpha \;=\; \sqrt{\frac{|K|\, d(x_0,x_1)^2}{N-1}}. \tag{20.13}$$

The values of β_t and its derivative β_t' (where the prime stands for partial derivative with respect to t) at the end-values $t=0$, $t=1$ will play a crucial role. It is a general fact that $\beta_1 = 1$, $\beta_0' = 0$. Let us set

$$\beta = \beta_0, \qquad \beta' = \beta_1'.$$

By explicit calculation,

$$\beta = \begin{cases} \left(\frac{\alpha}{\sin\alpha}\right)^{N-1} > 1 \\ 1 \\ \left(\frac{\alpha}{\sinh\alpha}\right)^{N-1} < 1 \end{cases} \qquad \beta' = \begin{cases} -(N-1)\left(1 - \frac{\alpha}{\tan\alpha}\right) < 0 & \text{if } K>0 \text{ and } \alpha \in [0,\pi], \\ 0 & \text{if } K = 0, \\ (N-1)\left(\frac{\alpha}{\tanh\alpha} - 1\right) > 0 & \text{if } K<0. \end{cases} \tag{20.14}$$

A standard Taylor expansion shows that, as $\alpha \to 0$ while K is fixed (which means that either $d(x_0,x_1) \to 0$ or $N \to \infty$), then

$$\beta \simeq 1 - \frac{K}{6}\, d(x_0,x_1)^2, \qquad \beta' \simeq -\frac{K}{3}\, d(x_0,x_1)^2,$$

whatever the sign of K.

The goal of this section is the following theorem:

Theorem 20.11 (distorted HWI inequality). *Let M be a Riemannian manifold equipped with a reference measure $\nu = e^{-V} \mathrm{vol}$, satisfying the curvature-dimension bound $\mathrm{CD}(K,N)$, $1 < N < \infty$. Let $U \in \mathcal{DC}_N$, and $p(r) = rU'(r) - U(r)$. Let $\mu_0 = \rho_0 \nu$ and $\mu_1 = \rho_1 \nu$ be two probability measures in $P_2^{\mathrm{ac}}(M)$, absolutely continuous with respect to ν. Further assume that $U(\rho_1)$ is integrable, ρ_1 has compact support, ρ_0 is Lipschitz continuous with compact support, $U(\rho_0)$ and $\rho_0 U'(\rho_0)$ are ν-integrable, and U' is bounded and Lipschitz on $\rho_0(M)$. Then*

$$\int_M U(\rho_0)\, d\nu \le \int_{M\times M} \left[U\left(\frac{\rho_1(x_1)}{\beta(x_0,x_1)}\right) \beta(x_0,x_1) \;+\; p(\rho_0(x_0))\beta'(x_0,x_1) \right.$$
$$\left. +\; U''(\rho_0(x_0))\, |\nabla^- \rho_0|(x_0)\, d(x_0,x_1) \right] \pi(dx_0\, dx_1), \quad (20.15)$$

where $\pi(dx_0\, dx_1)$ is the (unique) optimal coupling of (μ_0,μ_1), and the coefficients β and β' are defined in (20.14).

Remark 20.12. Recall from Remark 20.2 that the assumption of U' being bounded Lipschitz on $\rho_0(M)$ means that either ρ_0 is twice differentiable at the origin, or ρ_0 is bounded away from 0. If none of these assumptions is satisfied, it might be a good idea to regularize U into a smooth approximation U_ϵ that still lies in $\mathcal{DC}_N$, and then let ϵ go to 0. Such a strategy does not work well at the level of Theorem 20.6, because λ_{U_ϵ} might be very different from λ_U; but in the present case, the value of λ_U does not play any role. The main case of interest is

$$U(r) = U_N(r) = -Nr(r^{-\frac{1}{N}} - 1), \qquad U_\epsilon(r) = U_{N,\epsilon}(r) = -Nr\big[(r+\epsilon)^{-\frac{1}{N}} - 1\big].$$

With such a strategy, one can prove for instance the validity of inequality (20.15) when $U = U_N$, and ρ_0, ρ_1 are Lipschitz and compactly supported. For that it is sufficient to write (20.15) with $U = U_{N,\epsilon}$ and then pass to the limit in the various terms as $\epsilon \to 0$, using the monotone and dominated convergence theorems.

Proof of Theorem 20.11. Let $(\mu_t = \rho_t\, \nu)_{0\le t\le 1}$ be the unique Wasserstein geodesic joining μ_0 to μ_1. Recall from Theorem 17.28 the convexity inequality

$$\int U(\rho_t)\, d\nu \le (1-t) \int U\left(\frac{\rho_0}{\beta_{1-t}}\right) \beta_{1-t}\, d\pi + t \int U\left(\frac{\rho_1}{\beta_t}\right) \beta_t\, d\pi,$$

and transform this into

$$\int U\left(\frac{\rho_0}{\beta_{1-t}}\right) \beta_{1-t}\, d\pi \le \int U\left(\frac{\rho_1}{\beta_t}\right) \beta_t\, d\pi$$
$$+ \int \left[\frac{U\left(\frac{\rho_0}{\beta_{1-t}}\right) \beta_{1-t} - U(\rho_0)}{t} \right] d\pi - \frac{1}{t} \int \big[U(\rho_t) - U(\rho_0)\big]\, d\pi \quad (20.16)$$

(with obvious notation). The point is to pass to the limit in each of the terms of (20.16). This is quite easy to do formally, but one has to be a bit careful about integrability issues. (Skip all the rest of the proof at first reading.)

First term: First consider the left-hand side of (20.16).

If $K = 0$, there is nothing to prove.

If $K > 0$, then β_t is a decreasing function of t; since $U(r)/r$ is an increasing function of r, it follows that

$$U(\rho_0) \geq U\left(\frac{\rho_0}{\beta_{1-t}}\right)\beta_{1-t} \geq U\left(\frac{\rho_0}{\beta}\right)\beta, \tag{20.17}$$

and $U(\rho_0/\beta_{1-t})\beta_{1-t}$ converges monotonically to $U(\rho_0)$ as $t \to 0$.

Here there are two cases: either $\inf \rho_0 = 0$, but then U' has to be bounded close to 0, so $U(\rho_0/\beta)\beta$ should be bounded below by $-C\rho_0$ for some constant C; or $\inf \rho_0 > 0$, but then $\inf(\rho_0/\beta) > 0$ also, and again $U(\rho_0/\beta)\beta$ has to be bounded below by $-C\rho_0$. In all cases, $U(\rho_0/\beta)\beta$ is bounded below by an integrable function. This makes it possible to apply the monotone convergence theorem in (20.17) and conclude that

$$\int U\left(\frac{\rho_0}{\beta_{1-t}}\right)\beta_{1-t}\, d\pi \xrightarrow[t\to 0]{} \int U(\rho_0)\, d\pi.$$

Finally, if $K < 0$, then all inequalities are reversed:

$$U(\rho_0) \leq U\left(\frac{\rho_0}{\beta_{1-t}}\right)\beta_{1-t} \leq U\left(\frac{\rho_0}{\beta}\right)\beta,$$

and now it is sufficient to check that $U(\rho_0/\beta)\beta$ is bounded *above* by an integrable function. By assumption, ρ_0 and ρ_1 have compact support, so the function β remains bounded from above and below; and ρ_0 is bounded, so $U(\rho_0/\beta)\beta$ is bounded above and compactly supported, in particular it is integrable. Just as before, one can conclude that

$$\int U\left(\frac{\rho_0}{\beta_{1-t}}\right)\beta_{1-t}\, d\pi \xrightarrow[t\to 0]{} \int U(\rho_0)\, d\pi.$$

Second term: Now consider the term in $U(\rho_1/\beta_t)\beta_t$. Then the discussion is quite similar to the previous one. For instance, in the case of positive curvature,

$$U(\rho_1) \geq U\left(\frac{\rho_1}{\beta_t}\right)\beta_t \geq U\left(\frac{\rho_1}{\beta}\right)\beta,$$

and $U(\rho_1/\beta_t)\beta_t$ converges monotonically to $U(\rho_1/\beta)\beta$ as $t \downarrow 0$. To apply the monotone convergence theorem, it is sufficient to know that $U(\rho_1)$ is integrable, which was part of the assumptions. In all the cases, one has

$$\int U\left(\frac{\rho_1}{\beta_t}\right)\beta_t\, d\pi \xrightarrow[t\to 0]{} \int U\left(\frac{\rho_1}{\beta}\right)\beta\, d\pi.$$

Third term: By convexity of U, the function $b \longmapsto U(r/b)b$ is convex, with derivative $-p(r/b)$; so

$$U(\rho_0) - U\left(\frac{\rho_0}{\beta_{1-t}}\right)\beta_{1-t} \geq\ -p\left(\frac{\rho_0}{\beta_{1-t}}\right)(1-\beta_{1-t});$$

or, which is the same,

$$\frac{U\left(\frac{\rho_0}{\beta_{1-t}}\right)\beta_{1-t} - U(\rho_0)}{t} \leq\ p\left(\frac{\rho_0}{\beta_{1-t}}\right)\left(\frac{1-\beta_{1-t}}{t}\right). \tag{20.18}$$

Since U belongs to $\mathcal{DC}_N$, p is always nondecreasing. If $K > 0$, then the ratio $(1-\beta_{1-t})/t$ decreases as $t \downarrow 0$, while it increases if $K < 0$. In all cases, the right-hand side of (20.18) converges monotonically to $p(\rho_0)\beta'$. In the case of negative curvature, this is sufficient

to apply the monotone convergence theorem (because the left-hand side of (20.18) is integrable, say for $t = 1$). In the case of positive curvature, one should check that the right-hand side of (20.18) is bounded above by an integrable function; but this is obvious since this is a nonpositive expression. All in all,

$$\limsup_{t \downarrow 0} \int \frac{\left[U\left(\frac{\rho_0}{\beta_{1-t}}\right) \beta_{1-t} - U(\rho_0) \right]}{t} \, d\pi \leq \int p\left(\frac{\rho_0}{\beta_{1-t}}\right) \left(\frac{1 - \beta_{1-t}}{t}\right) \, d\pi.$$

Fourth term: By Theorem 20.1,

$$\limsup_{t \downarrow 0} \left(-\frac{1}{t} \int [U(\rho_t) - U(\rho_0)] \, d\pi \right) \leq \int U''(\rho_0(x_0)) \, |\nabla^- \rho_0|(x_0) \, d(x_0, x_1) \, \pi(dx_0 \, dx_1).$$

This concludes the proof of Theorem 20.11. □

Bibliographical Notes

Formula (20.6) appears as Theorem 5.30 in my book [365] when the space is $\mathbb{R}^n$; there were precursors, see for instance [290, 292]. The integration by parts leading from (20.6) to (20.7) is quite tricky because of regularity issues; I established it in $\mathbb{R}^n$ for some particular functions U, together with Cordero-Erausquin and Nazaret, in [119, Lemma 7], then the proof was extended to more general functions U (but still in $\mathbb{R}^n$) by Carrillo, McCann and myself [102, Lemma 5.12], for smooth densities. The same integration by parts formula was also established independently by Ambrosio, Gigli and Savaré [15, Lemma 10.4.5]. It is clear that the proof can be adapted to a Riemannian setting, however nobody has written it precisely, and I am not sure exactly what technical assumptions should be imposed. In any case, all of this is quite technical, which is why I preferred here to be content with Theorem 20.1. On one hand, the bound established in that theorem, although less precise than (20.6), is still sufficient for most applications known to me. On the other hand, its proof is quite simple, and applies to nonsmooth spaces, which will be quite useful in Part III of these notes. The argument is taken from my joint work with Lott [247] (where the space $\mathcal{X}$ is assumed to be compact, which simplifies a bit the assumptions).

The HWI inequality was established in my joint work with Otto [292]; obviously, it extends to any reasonable functional which is K-displacement convex. A precursor was studied by Otto [290]. An application to a "concrete" problem of partial differential equations can be found in [103, Section 5].

In the proof of Theorem 20.7, I assumed some regularity assumptions on the density ρ_0; but there are standard regularization procedures to remove them: First truncate ρ_0 close to 0 and ∞, then regularize the resulting density. In this way, one can build a family of probability measures ρ_0^n so that each ρ_0^n is smooth, bounded away from 0 and ∞, and $H_\nu(\rho_0^n \, \nu) \longrightarrow H_\nu(\rho_0 \, \nu)$, $I_\nu(\rho_0^n \, \nu) \longrightarrow I_\nu(\rho_0 \, \nu)$, $W_2(\rho_0^n \, \nu, \mu_1) \longrightarrow W_2(\rho_0 \, \nu, \, \mu_1)$. Closely related arguments can be found in [292, Appendix] and [365, Proof of Theorem 9.17, Step 1].

The role of the HWI inequality as an interpolation inequality is briefly discussed in [365, Section 9.4] and turned into application in [103, Proof of Theorem 5.1]: in that reference one studies rates of convergence for certain nonlinear partial differential equations, and combines a bound on the Fisher information with a convergence estimate in Wasserstein distance, to establish a convergence estimate in a stronger sense (L^1 norm, for instance).

A slightly different derivation of the HWI inequality is due to Cordero-Erausquin [115]; a completely different derivation is due to Bobkov, Gentil and Ledoux [56]. Also a slight variation of these inequalities was studied by Agueh, Ghoussoub and Kang [3].

The discussion for finite-dimensional inequalities is taken from joint work of Lott and myself.

21

Isoperimetric-type inequalities

It is a fact of experience that several inequalities with isoperimetric content can be retrieved by considering the above-tangent formulation of displacement convexity, when the final (target) measure is some reference measure. Here is a possible heuristic explanation for this phenomenon. Assume, for the sake of the discussion, that the initial measure is the normalized indicator function of some set A. Think of the functional U_ν as the internal energy of some fluid that is initially confined in A. In a displacement interpolation, some of the mass of the fluid will have to flow out of A, leading to a variation of the energy (typically, more space available means less density and less energy). The decrease of energy at initial time is related to the amount of mass that is able to flow out of A at initial time, and that in turn is related to the surface of A (a small surface leads to a small variation, because not much of the fluid can escape). So by controlling the decrease of energy, one should eventually get a control on the surface of A.

The functional nature of this approach makes it possible to replace the set A by some arbitrary probability measure $\mu = \rho\,\nu$. Then, what plays the role of the "surface" of A is some integral expression involving $\nabla\rho$. Any inequality expressing the domination of an integral expression of ρ by an integral expression of ρ and $\nabla\rho$ will be loosely referred to as a **Sobolev**-type, or **isoperimetric**-type inequality. Of course there are many many variants of such inequalities.

Logarithmic Sobolev inequalities

A probability measure ν on a Riemannian manifold is said to satisfy a logarithmic Sobolev inequality if the functional H_ν is dominated by (a constant multiple of) the functional I_ν. Here is a more precise definition:

Definition 21.1 (Logarithmic Sobolev inequality). *Let M be a Riemannian manifold, and ν a probability measure on M. It is said that ν satisfies a logarithmic Sobolev inequality with constant λ if, for any probability measure $\mu = \rho\,\nu$ with ρ Lipschitz, one has*

$$H_\nu(\mu) \leq \frac{1}{2\lambda} I_\nu(\mu). \tag{21.1}$$

Explicitly, inequality (21.1) means

$$\int \rho \log \rho \, d\nu \leq \frac{1}{2\lambda} \int \frac{|\nabla\rho|^2}{\rho} \, d\nu. \tag{21.2}$$

Equivalently, for any function u one should have

$$\int u^2 \, \log(u^2) \, d\nu - \left(\int u^2 \, d\nu \right) \log \left(\int u^2 \, d\nu \right) \leq \frac{2}{\lambda} \int |\nabla u|^2 \, d\nu. \tag{21.3}$$

To go from (21.2) to (21.3), just set $\rho = u^2 / (\int u^2 \, d\nu)$ and notice that $\nabla |u| \leq |\nabla u|$.

The Lipschitz regularity of ρ allows one to define $|\nabla \rho|$ pointwise, for instance by means of (20.1). Everywhere in this chapter, $|\nabla \rho|$ may also be replaced by the quantity $|\nabla^- \rho|$ appearing in (20.2); in fact both expressions coincide almost everywhere if u is Lipschitz.

This restriction of Lipschitz continuity is too strong, and can be relaxed with a bit of work. For instance, if $\nu = e^{-V} \, \mathrm{vol}$, with $V \in C^2(M)$, then one can use a little bit of distribution theory to show that the quantity $\int |\nabla \rho|^2 / \rho \, d\nu$ is well-defined in $[0, +\infty]$, and then (21.1) makes sense. But in the sequel, I shall just stick to Lipschitz functions. The same remark applies to other functional inequalities which will be encountered later: dimension-dependent Sobolev inequalities, Poincaré inequalities....

Logarithmic Sobolev inequalities are *dimension-free* Sobolev inequalities. Indeed, the dimension of the space does not appear explicitly in (21.3). This is one reason why these inequalities are extremely popular in various branches of statistical mechanics, mathematical statistics, quantum field theory, and more generally the study of phenomena in high or infinite dimension. They are also used in geometry and partial differential equations, including Perelman's recent work on the Ricci flow and the Poincaré conjecture.

At this stage of the course, the next theorem, a famous result in Riemannian geometry, will seem almost trivial.

Theorem 21.2 (Bakry–Émery theorem). *Let M be a Riemannian manifold equipped with a reference probability measure $\nu = e^{-V} \mathrm{vol}$, satisfying the curvature assumption $\mathrm{CD}(K, \infty)$ for some $K > 0$. Then ν satisfies a logarithmic Sobolev inequality with constant K, i.e.*

$$H_\nu \leq \frac{I_\nu}{2K}. \tag{21.4}$$

Example 21.3. For the Gaussian measure $\gamma(dx) = (2\pi)^{-n/2} e^{-|x|^2/2}$ in $\mathbb{R}^n$, one has

$$H_\gamma \leq \frac{I_\gamma}{2}, \tag{21.5}$$

independently of the dimension. This is the Stam–Gross logarithmic Sobolev inequality. By scaling, for any $K > 0$ the measure $\gamma_K(dx) = (2\pi/K)^{-n/2} e^{-K|x|^2} \, dx$ satisfies a logarithmic Sobolev inequality with constant K.

Remark 21.4. More generally, if $V \in C^2(\mathbb{R}^n)$, $\nabla^2 V \geq K \, I_n$, then Theorem 21.2 shows that $\nu(dx) = e^{-V(x)} \, dx$ satisfies a logarithmic Sobolev inequality with constant K. Example 21.3 shows that the constant K in (21.4) is optimal in general.

Remark 21.5. The curvature assumption $\mathrm{CD}(K, \infty)$ is quite restrictive, however there are known perturbation theorems which immediately extend the range of application of Theorem 21.2. For instance, if ν satisfies a logarithmic Sobolev inequality, v is a bounded function and $\widetilde{\nu} = e^{-v} \nu / Z$ is another probability measure obtained from ν by multiplication by e^{-v}, then also $\widetilde{\nu}$ satisfies a logarithmic Sobolev inequality (Holley–Stroock perturbation theorem). The same is true if v is unbounded, but satisfies $\int e^{\alpha |\nabla v|^2} \, d\nu < \infty$ for α large enough.

Proof of Theorem 21.2. From Corollary 19.13, ν admits square-exponential moments, in particular it lies in $P_2(M)$. Then from (20.10) and the inequality $ab \leq K a^2/2 + b^2/(2K)$,

$$H_\nu(\mu) \le W_2(\mu,\nu)\,\sqrt{I_\nu(\mu)} - \frac{K W_2(\mu,\nu)^2}{2} \;\le\; \frac{I_\nu(\mu)}{2K}.$$

□

Open Problem 21.6. *It is known that if M satisfies $\mathrm{CD}(K,N)$ with $N<\infty$, then the optimal constant in the Sobolev inequality is not K but $KN/(N-1)$. Can this be proven by a transport argument?*

In the next section, some finite-dimension Sobolev inequalities will be addressed, but it is not clear at all that they are strong enough to lead to the solution of Problem 21.6. Before examining these issues, I shall state an easy variation of Theorem 21.2:

Theorem 21.7 (Sobolev-L^∞ interpolation inequalities). *Let M be a Riemannian manifold, equipped with a reference probability measure $\nu = e^{-V}\mathrm{vol}$, $V \in C^2(M)$, satisfying the curvature-dimension condition $\mathrm{CD}(K,N)$ for some $K>0$, $N \in (1,\infty]$. Let further $U \in \mathcal{DC}_N$. Then, for any Lipschitz-continuous probability density ρ, and $\mu = \rho\,\nu$, one has the inequality*

$$U_\nu(\mu) - U_\nu(\nu) \le \frac{\max(\rho)^{\frac{1}{N}}}{2K}\, I_{U,\nu}(\mu), \tag{21.6}$$

where ρ is the density of μ with respect to ν, and

$$I_{U,\nu}(\mu) = \int_M \rho\,|\nabla U'(\rho)|^2\,d\nu. \tag{21.7}$$

Proof. The proof is the same as for Theorem 21.2, using the generalized HWI inequality of Theorem 20.10. Note that, by Jensen's inequality, $U_\nu(\mu) = \int U(\rho)\,d\nu \ge U(\int \rho\,d\nu) = U(1) = U_\nu(\nu)$. □

Sobolev inequalities

Sobolev inequalities are one among several classes of functional inequalities with isoperimetric content; they are extremely popular in the theory of partial differential equations theory. They look like logarithmic Sobolev inequalities, but with powers instead of logarithms, and they take dimension into account explicitly.

The most basic Sobolev inequality is in Euclidean space: If u is a function on $\mathbb{R}^n$ such that $\nabla u \in L^p(\mathbb{R}^n)$ $(1 \le p < n)$ and u vanishes at infinity (in whatever sense, see e.g. Remark 21.12 below), then u automatically lies in $L^{p^\star}(\mathbb{R}^n)$ where $p^\star = (np)/(n-p) > p$. More quantitatively, there is a constant $S = S(n,p)$ such that

$$\|u\|_{L^{p^\star}(\mathbb{R}^n)} \le S\,\|\nabla u\|_{L^p(\mathbb{R}^n)}.$$

There are other versions for $p=n$ (in which case essentially $\exp(cu^{n'})$ is integrable, $n' = n/(n-1)$), and $p>n$ (in which case u is Hölder-continuous). There are also many many variants for a function u defined on a set Ω that might be a reasonable open subset of either $\mathbb{R}^n$ or a Riemannian manifold M. For instance,

$$\|u\|_{L^{p^\star}(\Omega)} \le A\|\nabla u\|_{L^p(\Omega)} + C\|u\|_{L^{p^\sharp}(\partial\Omega)}, \qquad p^\sharp = \frac{(n-1)\,p}{n-p}, \quad 1 \le p < n,$$

$$\|u\|_{L^{p^\star}(\Omega)} \le A\|\nabla u\|_{L^p(\Omega)} + B\|u\|_{L^q(\Omega)} \qquad 1 \le p < n, \quad 1 \le q,$$

etc. One can also quote the **Gagliardo–Nirenberg** interpolation inequalities, which typically take the form

$$\|u\|_{L^{p^\star}} \le G\, \|\nabla u\|_{L^p}^{1-\theta}\, \|u\|_{L^q}^{\theta}, \qquad 1 \le p < n, \quad 1 \le q < p^\star, \qquad 0 \le \theta \le 1,$$

with some restrictions on the exponents. I will not say more about Sobolev-type inequalities, but there are entire books devoted to them.

In a Riemannian setting, there is a famous family of Sobolev inequalities obtained from the curvature-dimension bound $\mathrm{CD}(K,N)$ with $K > 0$ and $2 < N < \infty$:

$$\frac{c}{q-2}\left[\left(\int |u|^q\, d\nu\right)^{\frac{2}{q}} - \int |u|^2\, d\nu\right] \le \int |\nabla u|^2\, d\nu, \qquad 1 \le q \le \frac{2N}{N-2}, \quad c = \frac{NK}{N-1}. \tag{21.8}$$

When $q \to 2$, (21.8) reduces to Bakry–Émery's logarithmic Sobolev inequality. The most interesting member of the family is obtained when q is the critical exponent $2^\star = (2N)/(N-2)$, and then (21.8) becomes

$$\|u\|^2_{L^{\frac{2N}{N-2}}(M)} \le \|u\|^2_{L^2(M)} + \left(\frac{4}{N-2}\right)\left(\frac{N-1}{KN}\right) \|\nabla u\|^2_{L^2(M)}. \tag{21.9}$$

There is no loss of generality in assuming $u \ge 0$, since the inequality for general u follows easily from the inequality for nonnegative u. Let us then change unknowns by choosing $\rho = u^{2N/(N-2)}$. By homogeneity, it is also no loss of generality to assume that $\mu := \rho\,\nu$ is a probability measure. Then inequality (21.9) becomes

$$H_{N/2,\nu}(\mu) = -\frac{N}{2}\int (\rho^{1-\frac{2}{N}} - \rho)\, d\nu \le \frac{1}{2K}\int \frac{|\nabla\rho|^2}{\rho}\, \left(\frac{(N-1)(N-2)}{N^2}\rho^{-\frac{2}{N}}\right) d\nu. \tag{21.10}$$

The way in which I have written inequality (21.10) might look strange, but it has the merit to show very clearly how the limit $N \to \infty$ leads to the logarithmic Sobolev inequality $H_{\infty,\nu}(\mu) \le \int (|\nabla\rho|^2/\rho)\, d\nu$.

I don't know whether (21.10), or more generally (21.8), can be obtained by transport. Instead, I shall derive related inequalities, whose relation to (21.10) is still unclear. I also mention that (21.9) implies (21.6) in the case when $U = U_N$, as a consequence of the inequality

$$H_{N,\nu} \le \left(\frac{N-2}{N-1}\right) H_{N/2,\nu}.$$

Theorem 21.8 (Sobolev inequalities from $\mathrm{CD}(K,N)$). *Let M be a Riemannian manifold, equipped with a reference measure $\nu = e^{-V}\mathrm{vol}$, satisfying the curvature-dimension inequality $\mathrm{CD}(K,N)$ for some $K > 0$, $1 < N < \infty$. Then, for any probability density ρ, Lipschitz continuous and strictly positive, and $\mu = \rho\,\nu$, one has*

$$H_{N,\nu}(\mu) = -N\int_M (\rho^{1-\frac{1}{N}} - \rho)\, d\nu \le \int_M \Theta^{(N,K)}(\rho, |\nabla\rho|)\, d\nu, \tag{21.11}$$

where

$$\Theta^{(N,K)}(r,g) = r \sup_{0\le\alpha\le\pi}\left(\frac{N-1}{N}\frac{g}{r^{1+\frac{1}{N}}}\sqrt{\frac{N-1}{K}}\,\alpha + N\left(1 - \left(\frac{\alpha}{\sin\alpha}\right)^{1-\frac{1}{N}}\right) + (N-1)\left(\frac{\alpha}{\tan\alpha} - 1\right) r^{-\frac{1}{N}}\right). \tag{21.12}$$

As a consequence,

$$H_{N,\nu}(\mu) \le \frac{1}{2K} \int_M \frac{|\nabla\rho|^2}{\rho} \left(\left(\frac{N-1}{N} \right)^2 \frac{\rho^{-\frac{2}{N}}}{\frac{1}{3} + \frac{2}{3}\rho^{-\frac{1}{N}}} \right) d\nu. \tag{21.13}$$

Remark 21.9. By taking the limit as $N \to \infty$ in (21.13), one recovers again the logarithmic Sobolev inequality of Bakry and Émery, with the sharp constant. For fixed N, the exponents appearing in (21.13) are sharp: For large ρ, the integrand in the right-hand side behaves like $|\nabla\rho|^2 \rho^{-(1+2/N)} = c_N |\nabla \rho^{\frac{1}{2^\star}}|^2$, so the critical Sobolev exponent 2^* governs this inequality, which is good. On the other hand, the *constants* appearing in (21.13) are definitely not sharp; for instance it is obvious that they do not imply exponential integrability as $N \to 2$.

Open Problems 21.10. *Is inequality* (21.11) *stronger, weaker, or not comparable to inequality* (21.10)*? Does inequality* (21.13) *follow from* (21.10)*? Can one find a transport argument leading to* (21.10)*?*

Proof of Theorem 21.8. Start from Theorem 20.11 and choose $U(r) = -N(r^{1-\frac{1}{N}} - r)$. After some straightforward calculations, it follows that

$$H_{N,\nu}(\mu) \le \int_M \theta^{(N,K)}\big(\rho, |\nabla\rho|, \alpha\big),$$

where $\alpha = \sqrt{K/(N-1)}\, d(x_0, x_1) \in [0, \pi]$, and $\theta^{(N,K)}$ is an explicit function such that

$$\Theta^{(N,K)}(r, g) = \sup_{\alpha \in [0,\pi]} \theta^{(N,K)}(r, g, \alpha).$$

This is sufficient to prove (21.11).

To go from (21.11) to (21.13), one can use the elementary inequalities

$$\begin{cases} N \left(1 - \left(\dfrac{\alpha}{\sin\alpha} \right)^{1-\frac{1}{N}} \right) \le - \left(\dfrac{N-1}{6} \right) \alpha^2 \\ (N-1) \left(\dfrac{\alpha}{\tan\alpha} - 1 \right) \le - \left(\dfrac{N-1}{3} \right) \alpha^2. \end{cases} \tag{21.14}$$

(See the bibliographical notes for proofs.) Plug these inequalities into (21.12), then compute the supremum explicitly: This gives the desired estimate of $\Theta^{(N,K)}$ from above. □

Now I shall consider the case of the Euclidean space $\mathbb{R}^n$, equipped with the Lebesgue measure, and show that sharp Sobolev inequalities can be obtained by a transport approach. The proof will take advantage of the scaling properties in $\mathbb{R}^n$.

Theorem 21.11 (Sobolev inequalities in $\mathbb{R}^n$). *Whenever u is a Lipschitz, compactly supported function on $\mathbb{R}^n$, then*

$$\|u\|_{L^{p^\star}(\mathbb{R}^n)} \le S_n(p)\ \|\nabla u\|_{L^p(\mathbb{R}^n)} \qquad 1 \le p < n, \quad p^\star = \frac{np}{n-p}, \tag{21.15}$$

where the constant $S_n(p)$ is given by

$$S_n(p) = \inf \left\{ \frac{p\,(n-1)}{n\,(n-p)}\ \frac{\left(\int |g| \right)^{\frac{1}{p^\star}} \left(\int |y|^{p'} |g(y)|\, dy \right)^{\frac{1}{p'}}}{\int |g|^{1-\frac{1}{n}}} \right\}, \qquad p' = \frac{p}{p-1},$$

and the infimum is taken over all functions $g \in L^1(\mathbb{R}^n)$, not identically 0.

be

Remark 21.12. The assumption of Lipschitz continuity for u can be removed, but I shall not do so here. Actually, inequality (21.15) holds true as soon as u is locally integrable and vanishes at infinity, in the sense that the Lebesgue measure of any set of the form $\{|u| \geq r\}$ $(r > 0)$ is finite.

Remark 21.13. The constant $S_n(p)$ is optimal.

Proof of Theorem 21.11. Choose $M = \mathbb{R}^n$, $\nu =$ Lebesgue measure, and apply Theorem 20.11 with $K = 0$, $N = n$ (recall Remark 20.12), and $\mu_0 = \rho_0\,\nu$, $\mu_1 = \rho_1\,\nu$, both of them compactly supported. Here $\beta \equiv 1$, $\beta' \equiv 0$, so formula (20.15) simplifies into

$$H_{n,\nu}(\mu_0) - H_{n,\nu}(\mu_1) \leq \left(1 - \frac{1}{n}\right) \int_{\mathbb{R}^n \times \mathbb{R}^n} \rho_0(x_0)^{-(1+\frac{1}{n})}\, |\nabla \rho_0|(x_0)\, d(x_0, x_1)\, \pi(dx_0\, dx_1).$$

By Hölder's inequality and the marginal property of π,

$$H_{n,\nu}(\mu_0) - H_{n,\nu}(\mu_1) \leq \left(1 - \frac{1}{n}\right) \left(\int_{\mathbb{R}^n} \rho_0^{-p(1+\frac{1}{n})}\, |\nabla \rho_0|^p\, d\mu_0 \right)^{\frac{1}{p}} \left(\int_{\mathbb{R}^n \times \mathbb{R}^n} d(x_0, x_1)^{p'}\, \pi(dx_0\, dx_1) \right)^{\frac{1}{p'}},$$

where $p' = p/(p-1)$. This can be rewritten

$$n \int \rho_1^{1-\frac{1}{n}}\, d\nu \leq n \int \rho_0^{1-\frac{1}{n}}\, d\nu + \left(1 - \frac{1}{n}\right) \left(\rho_0^{-p(1+\frac{1}{n})}\, |\nabla \rho_0|^p\, d\mu_0 \right)^{\frac{1}{p}} W_{p'}(\mu_0, \mu_1). \quad (21.16)$$

Now I shall use a homogeneity argument. Fix ρ_1 and ρ_0 as above, and define $\rho_0^{(\lambda)}(x) = \lambda^n \rho_0(\lambda x)$. On one hand,

$$\int \left(\rho_0^{(\lambda)}\right)^{1-\frac{1}{n}}\, d\nu = \lambda^{-1} \int \rho_0^{1-\frac{1}{n}}\, d\nu \xrightarrow[\lambda \to \infty]{} 0;$$

on the other hand,

$$\int_{\mathbb{R}^n} (\rho_0^{(\lambda)})^{-p(1+\frac{1}{n})}\, |\nabla \rho_0^{(\lambda)}|^p\, d\mu_0^{(\lambda)} \qquad \text{does not depend on } \lambda.$$

Moreover, as $\lambda \to \infty$, the probability measure $\mu_0^{(\lambda)} = \rho_0^{(\lambda)}\,\nu$ converges weakly to the Dirac mass δ_0 at the origin; so

$$W_{p'}(\mu_0^{(\lambda)}, \mu_1) \longrightarrow W_{p'}(\delta_0, \mu_1) = \left(\int |y|^{p'}\, d\mu_1(y) \right)^{\frac{1}{p'}}.$$

So after writing (21.16) for $\mu_0 = \mu_0^{(\lambda)}$ and then passing to the limit as $\lambda \to \infty$, one obtains

$$n \int \rho_1^{1-\frac{1}{n}}\, d\nu \leq \left(\int_{\mathbb{R}^n} \rho_0^{-p(1+\frac{1}{n})}\, |\nabla \rho_0|^p\, d\mu_0 \right)^{\frac{1}{p}} \left(\int |y|^{p'}\, d\mu_1(y) \right)^{\frac{1}{p'}}. \quad (21.17)$$

Let us change unknowns and define $\rho_0 = u^{1/p^\star}$, $\rho_1 = g$; then (21.17) becomes

$$1 \leq \frac{p\,(n-1)}{n\,(n-p)} \left(\frac{\left(\int |y|^{p'}\, g(y)\, dy \right)^{\frac{1}{p'}}}{\int g^{p^\star(1-\frac{1}{n})}} \right) \|\nabla u\|_{L^p},$$

where u and g are only required to satisfy $\int u^{p^\star} = 1$, $\int g = 1$. Then (21.15) follows by homogeneity again. □

To conclude this section, I shall consider the case when $\mathrm{Ric}_{N,\nu} \geq K < 0$ and derive Sobolev inequalities for compactly supported functions. Since I shall not be concerned here with optimal constants, I shall only discuss the limit case $p = 1$, $p^\star = n/(n-1)$, which implies the general inequality for $p < n$ (via Hölder's inequality), up to a loss in the constants.

Theorem 21.14 (L^1 Sobolev inequalities from $\mathrm{CD}(K,N)$). *Let M be a Riemannian manifold equipped with a reference measure ν, satisfying the curvature-dimension bound $\mathrm{CD}(K,N)$ for some $K < 0$, $N \in (1,\infty)$. Then, for any ball $\mathcal{B} = B(z,R)$, $R \geq 1$, there are constants A and B, only depending on a lower bound on K, and upper bounds on N and R, such that for any Lipschitz function u supported in $\mathcal{B}$,*

$$\|u\|_{L^{\frac{N}{N-1}}} \leq A\,\|\nabla u\|_{L^1} + B\,\|u\|_{L^1}. \tag{21.18}$$

Proof of Theorem 21.14. Inequality (21.18) remains unchanged if we multiply ν by a positive constant. So we might assume, without loss of generality, that $\nu[B(z,R)] = 1$.

Apply again (20.15), in the form

$$N - \int \rho_0^{1-\frac{1}{N}}\,d\nu \leq N - \int \rho_1^{1-\frac{1}{N}}\,\beta^{\frac{1}{N}}\,d\pi + \int \rho_0^{1-\frac{1}{N}}\,\beta'\,d\pi + \frac{1}{N}\int \rho_0^{-\frac{1}{N}}\,|\nabla\rho_0|\,d(x_0,x_1)\,d\pi, \tag{21.19}$$

and choose $\rho_1 = 1_{B(x_0,R)}/\nu[B(x_0,R)]$ (the normalized indicator function of the ball). The arguments of β and β' in (21.19) belong to $B(z,R)$, so the coefficients β and β' remain bounded by some explicit function of N, K and R; and the distance $d(x_0,x_1)$ remains bounded by $2R$. So there are constants $\delta(K,N,R) > 0$ and $\overline{C}(K,N,R)$ such that

$$-\int \rho_0^{1-\frac{1}{N}} \leq -\,\delta(K,N,R)\,\nu[\mathcal{B}]^{\frac{1}{N}} + \overline{C}(K,N,R)\left[\int \rho_0^{1-\frac{1}{N}} + \int \rho_0^{-\frac{1}{N}}|\nabla\rho_0|\right]. \tag{21.20}$$

Recall that $\nu[\mathcal{B}] = 1$. Then after the change of unknowns $\rho_0 = u^{N/(N-1)}$, inequality (21.20) implies

$$1 \leq S\big(K,N,R\big)\left[\|\nabla u\|_{L^1(M)} + \|u\|_{L^1(M)}\right],$$

for some explicit constant $S = (\overline{C}+1)/\delta$. This holds true under the constraint $1 = \int \rho = \int u^{N/(N-1)}$, and then inequality (21.18) follows by homogeneity. □

Isoperimetric inequalities

Isoperimetric inequalities can sometimes be obtained as limits of Sobolev inequalities applied to indicator functions. The most classical example is the equivalence between the optimal Sobolev inequality $\|u\|_{L^{n/(n-1)}(\mathbb{R}^n)} \leq S_1(1)\,\|\nabla u\|_{L^1(\mathbb{R}^n)}$ and the Euclidean isoperimetric inequality

$$\frac{|\partial A|}{|A|^{\frac{n-1}{n}}} \geq \frac{|\partial B^n|}{|B^n|^{\frac{n-1}{n}}}$$

considered in Chapter 2.

As seen before, there is a proof of the optimal Sobolev inequality in $\mathbb{R}^n$ based on transport, and of course this leads to a proof of the Euclidean isoperimetry. There is also a more direct path to a transport-based proof of isoperimetry, as explained in Chapter 2.

Apart from the Euclidean one, the most famous isoperimetric inequality in differential geometry is certainly the **Lévy–Gromov inequality**, which states that if A is a reasonable set in a manifold (M,g) with dimension n and Ricci curvature bounded below by K, then

$$\frac{|\partial A|}{|A|^{\frac{n-1}{n}}} \geq \frac{|\partial B|}{|B|^{\frac{n-1}{n}}},$$

where B is a spherical cap in the model sphere S (that is, the sphere with dimension N and Ricci curvature K) such that $|B|/|S| = |A|/|M|$. In other words, isoperimetry in M is at least as strong as isoperimetry in the model sphere.

I don't know if the Lévy–Gromov inequality can be retrieved from optimal transport, and I think this is one of the most exciting open problems in the field. Indeed, there is to my knowledge no "reasonable" proof of the Lévy–Gromov inequality, in the sense that the only known arguments rely on subtle results from geometric measure theory, about the rectifiability of certain extremal sets. A softer argument would be conceptually very satisfactory. I record this in the form of a loosely formulated open problem:

Open Problem 21.15. *Find a transport-based, soft proof of the Lévy–Gromov isoperimetric inequality.*

The same question can be asked for the Gaussian isoperimetry, which is the infinite-dimensional version of the Lévy–Gromov inequality. In that case however there are known functional versions, and softer approaches.

Poincaré inequalities

Poincaré inequalities are related to Sobolev inequalities, and often appear as limit cases of them. (I am sorry if the reader begins to be bored by this litany: Logarithmic Sobolev inequalities are limits of Sobolev inequalities, isoperimetric inequalities are limits of Sobolev inequalities, Poincaré inequalities are limits of Sobolev inequalities...) Here in this section I shall only consider *global* Poincaré inequalities, which are rather different from the local inequalities considered in Chapter 19.

Definition 21.16 (Poincaré inequalities). *Let M be a Riemannian manifold, and ν a probability measure on M. It is said that ν satisfies a Poincaré inequality with constant λ if, for any $u \in L^2(\mu)$ with u Lipschitz, one has*

$$\big\| u - \langle u \rangle \big\|^2_{L^2(\nu)} \leq \frac{1}{\lambda} \, \|\nabla u\|^2_{L^2(\nu)}, \qquad \langle u \rangle = \int u \, d\nu. \tag{21.21}$$

Inequality (21.21) can be reformulated into

$$\int u \, d\nu = 0 \implies \qquad \|u\|^2_{L^2} \leq \frac{\|\nabla u\|^2_{L^2}}{\lambda}.$$

This writing makes the formal connection with the logarithmic Sobolev inequality very natural. (The Poincaré inequality is obtained as the limit of the logarithmic Sobolev inequality when one sets $\mu = (1 + \varepsilon u)\, \nu$ and lets $\varepsilon \to 0$.)

Like Sobolev inequalities, Poincaré inequalities express the domination of a function by its gradient; but unlike Sobolev inequalities, they do not include any gain of integrability. Poincaré inequalities have spectral content, since the best constant λ can be interpreted as the spectral gap for the Laplace operator on M.[1] There is no Poincaré inequality on $\mathbb{R}^n$

[1] This is one reason to take λ as the constant defining the Poincaré inequality. Unfortunately this is not consistent with the convention that I used for local Poincaré inequalities; another choice would have been to call λ^{-1} the Poincaré constant.

equipped with the Lebesgue measure (the usual "flat" Laplace operator does not have a spectral gap), but there is a Poincaré inequality on, say, any compact Riemannian manifold.

Establishing Poincaré inequalities for various measures is an extremely classical problem on which a lot has been written. Here is one of the oldest results in the field:

Theorem 21.17 (Lichnérowicz's spectral gap inequality). *Let M be a Riemannian manifold equipped with a reference measure ν, satisfying the curvature-dimension condition $\mathrm{CD}(K,N)$ for some $K > 0$, $N \in (1, +\infty]$. Then, ν satisfies a Poincaré inequality with constant $KN/(N-1)$.*

In other words, if $\mathrm{CD}(K,N)$ holds true, then for any Lipschitz function f on M with $\int f\, d\nu = 0$, one has

$$\left[\int f\, d\nu = 0\right] \Longrightarrow \qquad \int f^2\, d\nu \le \frac{N-1}{KN} \int |\nabla f|^2\, d\nu. \tag{21.22}$$

Remark 21.18. If $\nu = e^{-V}\,\mathrm{vol}$ and $L = \Delta - \nabla V \cdot \nabla$, $V \in C^2(M)$, then (21.22) means that L admits a spectral gap of size at least $KN/(N-1)$:

$$\lambda_1(-L) \ge \frac{KN}{N-1}.$$

Proof. In the case $N < \infty$, apply (21.13) with $\mu = (1 + \varepsilon f)\,\nu$, where ε is a small positive number, f is Lipschitz and $\int f\, d\nu = 0$. Since M has finite diameter, f is bounded, so μ is a probability measure for ε small enough. Then, by standard Taylor expansion of the logarithm function,

$$H_{N,\nu}(\mu) = \varepsilon \int f\, d\nu + \varepsilon^2 \left(\frac{N-1}{N} \int \frac{f^2}{2}\, d\nu\right) + o(\varepsilon^2),$$

and the first term on the right-hand side vanishes by assumption. Similarly,

$$\int \frac{|\nabla \rho|^2}{\rho} \left(\frac{\rho^{-\frac{2}{N}}}{\frac{1}{3} + \frac{2}{3}\rho^{-\frac{1}{N}}}\right) = \varepsilon^2 \int |\nabla f|^2\, d\nu + o(\varepsilon^2).$$

So (21.13) implies

$$\frac{N-1}{N} \int \frac{f^2}{2}\, d\nu \le \frac{1}{2K} \left(\frac{N-1}{N}\right)^2 \int |\nabla f|^2\, d\nu,$$

and then inequality (21.22) follows.

In the case $N = \infty$, start from inequality (21.4) and apply a similar reasoning (it is in fact a well-known property that a logarithmic Sobolev inequality with constant K implies a Poincaré inequality with constant K). □

Bibliographical Notes

Standard sources about classical isoperimetric inequalities are the book by Burago and Zalgaller [82], and the survey by Osserman [**?**]. A very general discussion of isoperimetric inequalities can be found in Bobkov and Houdré [58]. As part of his huge work on concentration of measure, Talagrand has put forward the use of isoperimetric inequalities in product spaces [346].

There are entire books devoted to logarithmic Sobolev inequalities; this subject goes back at least to Nelson [283] and Gross [198], in relation with hypercontractivity and quantum field theory; but it also takes its roots in earlier works by Stam [333] and Bonami [64]. A gentle introduction, and references, can be found in [21]. The 1992 survey by Gross [199], the Saint-Flour course by Bakry [30] and the book by Royer [**?**] are classical references. Applications to concentration theory and deviation inequalities can also be found in those sources, or in Ledoux's synthesis works [235, 237].

The first and most famous logarithmic Sobolev inequality is the one that holds true for the Gaussian reference measure in $\mathbb{R}^n$, (equation (21.5)). An "equivalent" formulation (in dimension 1) was derived by Stam [333], so I personally use the name of "Stam–Gross logarithmic Sobolev inequality for (21.5); of course, this is debatable. There are more than fifteen known proofs of this inequality; see Gross [199] for a partial list.

The Bakry–Émery theorem (Theorem 21.2 was proven in [31] by a semigroup method which will be reinterpreted in Chapter 25 as a gradient flow argument. The proof was rewritten in a language of partial differential equations in [23], with emphasis on the link to convergence to equilibrium for the heat-like equation $\partial_t \rho = L\rho$.

The proof of Theorem 21.2 given in these notes is essentially the one that appeared in my joint work with Otto [292]. When the manifold M is $\mathbb{R}^n$ (and V is K-convex), there is a slightly simpler variant of that argument, due to Cordero-Erausquin [115]; there are also two quite different proofs, one by Caffarelli [91] (based on Caffarelli's log concave perturbation theorem) and one by Bobkov and Ledoux [59] (based on the Brunn–Minkowski inequality in $\mathbb{R}^n$). It is likely that the distorted Prékopa–Leindler inequality (Theorem 19.12) can be used to derive an alternative proof of the Bakry–Émery theorem in the style of Bobkov–Ledoux.

The Holley–Stroock perturbation theorem for logarithmic Sobolev inequalities, explained in Remark [**?**], was proven in [212]. The other criterion mentioned in Remark [**?**] namely $\int e^{\alpha |\nabla v|^2} \, d\nu < \infty$ for α large enough, is due to Aida [**?**].

The refinement of the constant in the logarithmic Sobolev inequalities by a dimensional factor of $N/(N-1)$ is somewhat tricky; see for instance Ledoux [232]. As a limit case, on S^1 there is a logarithmic Sobolev inequality with constant 1, although the Ricci curvature vanishes identically on S^1.

Sobolev inequalities also fill up books, but usually the emphasis is more on regularity issues; in fact, for a long time logarithmic Sobolev inequalities and plain Sobolev inequalities were used and studied by quite different communities. A standard reference is the book by Maz'ja [264], but there are many alternative sources. A good synthetic source for the family (21.8) is the course by Ledoux [236]. In that reference the author shows how to deduce some geometric information from this family of inequalities.

Demange has recently obtained a derivation of (21.10) which is, from my point of view, very satisfactory, and will be explained later in Chapter 25. By Demange's method one can establish the following generalization of (21.10): under adequate regularity assumptions, if (M, ν) satisfies the curvature-dimension bound $\mathrm{CD}(K, N)$, and $U \in \mathcal{DC}_N$, and A is defined by $A(0) = 0$ and $A(1) = 0$, $A''(r) = r^{-1/N} U''(r)$, then for any probability density ρ,

$$\int_M A(\rho) \, d\nu \leq \frac{1}{2K} \int_M \rho^{1 - \frac{1}{N}} \left| \nabla U'(\rho) \right|^2 d\nu.$$

Many other variants, some of them rather odd-looking, appear in Demange's work [136, 137, 135]. For instance, he is able to establish apparently sharp inequalities for nonlinearities U satisfying the following condition:

$$\frac{d}{dr}\left[r\left(\frac{rU''(r)}{U'(r)}+\frac{1}{N}\right)\right] \geq \frac{9N}{4(N+2)}\left(\frac{rU''(r)}{U'(r)}+\frac{1}{N}\right)^2.$$

Demange also pointed out to me that (21.9) implies (21.6), without any loss in the constants. It is interesting to note that (21.6) can be proven very simply by a transport argument, while no such thing is known for (21.9).

The proof of Theorem 21.8 is taken from a collaboration with Lott [249]. Inequalities (21.14) are proven in Section 5 of this reference.

The use of transport methods to study isoperimetric inequalities in $\mathbb{R}^n$ goes back at least to Knothe [228]; Gromov [275, Appendix] revived the interest in Knothe's approach by using it to prove the isoperimetric inequality in $\mathbb{R}^n$. Recently, the method was put to a higher degree of sophistication by Cordero-Erausquin, Nazaret and myself [119]. In this work, we recover general optimal Sobolev inequalities in $\mathbb{R}^n$, together with some families of optimal Gagliardo–Nirenberg inequalities. (The proof of the Sobolev inequalities is reproduced in [365, Theorem 6.21].) The results themselves are not new, since optimal Sobolev inequalities in $\mathbb{R}^n$ were established independently by Aubin, Talenti and Rodemich, already in the seventies (see [119] for references), while the optimal Gagliardo–Nirenberg inequalities were discovered by Dolbeault and Del Pino [131]. However, I think that all in all the transport approach is simpler, especially for the Gagliardo–Nirenberg family. In [119] the optimal Sobolev inequalities came with a "dual" family of inequalities, that can be interpreted as a particular case of so-called Faber–Krahn inequalities; there is still (at least for me) some mystery in this duality.

An interesting feature of the proof by Cordero-Erausquin, Nazaret and myself (shared by Gromov's proof of isoperimetry) is the fact that it is insensitive to the choice of norm in $\mathbb{R}^n$. This remark will become interesting inthe las part of these notes.

In the present chapter, I have modified a bit the argument of [119] to avoid the use of the Alexandrov theorem about second derivatives of convex functions (Theorem 14.23 in the first Appendix of Chapter 14). The advantage is to get a simpler proof, however the computations are less precise, and some useful "magic" cancellations (such as $x + (\nabla\varphi - x) = \nabla\varphi$) are not available any longer; I used a homogeneity argument to get around this problem. A drawback of this approach is that the discussion about cases of equality is not possible any longer (anyway a clean discussion of equality cases requires much more efforts; see [119, Section 4]). The proof presented here should work through if $\mathbb{R}^n$ is replaced by a cone with nonnegative Ricci curvature, although I did not check details.

In the new argument, the effect of the homogeneity is to transform a given functional inequality, seemingly not optimal, into the optimal one. I wonder whether a similar argument could lead from (21.6) to (21.9) on the sphere.

After [119], Maggi and myself [251] pushed the method even further further, to recover "very optimal" Sobolev inequalities with trace terms, in $\mathbb{R}^n$. This settled some problems that had been left open in a classical work by Brézis and Lieb [80]. Much more information can be found in [251], and recently we wrote a sequel [252] in which limit cases (such as inequalities of Moser–Trudinger type) are considered.

As far as all these applications of transport to Sobolev or isoperimetric inequalities in $\mathbb{R}^n$ are concerned, the Knothe coupling works about just as fine as the optimal coupling. I am not completely sure that it can be used in subtle refinements such as the discussion of equality cases performed in [119], but I would not be surprised if the answer were affirmative. In any case, if the reader is looking for a transport argument related to some geometric inequality in $\mathbb{R}^n$, I personally advise him or her to try the Knother coupling first.

The Lévy–Gromov inequality was first conjectured by Lévy in the case when the manifold M is the boundary of a uniformly convex set (so the *sectional* curvatures are bounded below by a positive constant). Lévy thought he had a proof, but his argument was faulty and repaired by Gromov [196]. There have also been some striking works by Bobkov, Ledoux and Bakry on the infinite-dimensional version of the Lévy–Gromov inequality (often called Gaussian isoperimetry); for this inequality there is an elegant functional formulation [55, 33]. On that subject I also warmly recommend (as usual) the synthesis works by Ledoux [233, 234].

The Lichnérowicz spectral gap theorem is usually encountered as a simple application of the Bochner formula. The above proof of Theorem 21.17 is a variant of the one which appears in my joint work with Lott [249]. Although less simple than the classical proof, it has the advantage, for the purpose of these notes, to be based on optimal transport. This is actually, to my knowledge, the first time that the dimensional refinement in the constants by a factor $N/(N-1)$ in an "infinite-dimensional functional inequality" is obtained from a transport argument.

22

Concentration inequalities

The theory of concentration of measure is a set of ideas, tools and recipes built on the idea that if a set A is given in a metric probability space $(\mathcal{X}, d, \mathbb{P})$, then the enlargement $A^r := \{x;\ d(x,A) \le r\}$ might acquire a very high probability as r increases. There is an equivalent statement that Lipschitz functions $\mathcal{X} \to \mathbb{R}$ are "almost constant" in the sense that they have a very small probability to deviate from some typical quantity, for instance their mean value. This theory was founded by Lévy and later developed by many authors, in particular Milman, Gromov and Talagrand.

To understand the relation between the two sides of concentration (sets and functions), it is most natural to think in terms of **median**, rather than mean value. By definition, a real number m_f is a median of the random variable $f : \mathcal{X} \to \mathbb{R}$ if

$$\mathbb{P}[f \ge m_f] \ge \frac{1}{2}; \qquad \mathbb{P}[f \le m_f] \ge \frac{1}{2}.$$

Then the two statements

(a) $\forall A \subset \mathcal{X},\ \forall r \ge 0, \quad \mathbb{P}[A] \ge 1/2 \Longrightarrow\ \mathbb{P}[A^r] \ge 1 - \psi(r)$

(b) $\forall f \in \mathrm{Lip}(\mathcal{X}),\ \forall r \ge 0, \quad \mathbb{P}[f > m_f + r] \le \psi(r/\|f\|_{\mathrm{Lip}})$

are equivalent. Indeed, to pass from (a) to (b), first reduce to the case $\|f\|_{\mathrm{Lip}} = 1$ and let $A = \{f \le m_f\}$; to pass from (b) to (a), let $f = d(\cdot, A)$ and note that 0 is a median of f.

The typical and most emblematic example of concentration of measure occurs in the Gaussian probability space $(\mathbb{R}^n, \gamma)$:

$$\gamma[A] \ge \frac{1}{2} \Longrightarrow \quad \gamma[A^r] \ge 1 - e^{-\frac{r^2}{2}}.$$

Here is the translation in terms of Lipschitz functions: If X is a Gaussian random variable with law γ, then for all Lipschitz functions $f : \mathbb{R}^n \to \mathbb{R}$,

$$\mathbb{P}\big[f(X) \ge \mathbb{E}\, f(X) + r\big] \le \exp\left(-\frac{r^2}{2\,\|f\|_{\mathrm{Lip}}^2}\right).$$

Another famous example is the unit sphere S^N: if σ^N stands for the normalized volume on S^N, then the formulas above can be replaced by

$$\sigma^N[A] \ge \frac{1}{2} \Longrightarrow \quad \sigma^N[A^r] \ge 1 - e^{-\frac{(N-1)}{2} r^2},$$

$$\mathbb{P}\big[f(X) \ge \mathbb{E}\, f(X) + r\big] \le \exp\left(-\frac{(N-1)r^2}{2\,\|f\|_{\mathrm{Lip}}^2}\right).$$

On this example we see that the phenomenon of concentration of measure becomes more and more important as the dimension increases to infinity.

Optimal transport and concentration

As first understood by Marton, there is a simple and robust functional approach to concentration inequalities based on optimal transport. One can encode some information about the concentration of measure with respect to some reference measure ν, by functional inequalities of the form

$$\forall \mu \in P(\mathcal{X}), \quad C(\mu, \nu) \leq \mathcal{E}_\nu(\mu), \tag{22.1}$$

where $C(\mu, \nu)$ is the optimal transport cost between μ and ν, and $\mathcal{E}_\nu$ is some local nonlinear functional ("energy") of μ.

This principle may be heuristically understood as follows. To any given measurable set A, associate the conditional measure $\mu_A = (1_A/\nu[A])\,\nu$. If the measure of A is not too small, then the associated energy $\mathcal{E}_\nu(\mu_A)$ will not be too high, and by (22.1) the optimal transport cost $C(\mu_A, \nu)$ will not be too high either. In that sense, the whole space $\mathcal{X}$ can be considered as a "small enlargement" of just A.

Here is a fluid mechanics analogy: imagine μ as the density of a fluid. The term on the right-hand side of (22.1) measures how difficult it is to prepare μ, for instance to confine it within a set A (this has to do with the measure of A); while the term on the left-hand side says how difficult it is for the fluid to invade the whole space, after it has been prepared initially with density μ.

The most important class of functional inequalities of the type (22.1) occurs when the cost function is of the type $c(x, y) = d(x, y)^p$, and the "energy" functional is the square root of Boltzmann's H functional,

$$H_\nu(\mu) = \int \rho \log \rho \, d\nu, \quad \mu = \rho\, \nu,$$

with the understanding that $H_\nu(\mu) = +\infty$ if μ is not absolutely continuous with respect to ν. Here below is a precise definition of these functional inequalities:

Definition 22.1 (T_p inequality). *Let $(\mathcal{X}, d)$ be a Polish space and let $p \in [1, \infty)$. Let ν be a reference probability measure in $P_p(\mathcal{X})$, and let $\lambda > 0$. It is said that ν satisfies a T_p inequality with constant λ if*

$$\forall \mu \in P_p(\lambda), \qquad W_p(\mu, \nu) \leq \sqrt{\frac{2\, H_\nu(\mu)}{\lambda}}.$$

These inequalities are often called transportation-cost inequalities, or **Talagrand inequalities**, although the latter denomination is sometimes restricted to the case $p = 2$.

Remark 22.2. Since $W_p \leq W_q$ for $p \leq q$, the T_p inequalities are stronger and stronger when p increases. The inequalities T_1 and T_2 have deserved most attention. It is an experimental fact that T_1 is more handy and flexible, while T_2 has more geometric content, and behaves better in large dimension (see for instance Corollary 22.6 below).

There are two important facts to know about T_p inequalities when p varies in the range $[1, 2]$: they admit a *dual formulation*, and they *tensorize*. These properties are described in the two Propositions below.

Proposition 22.3 (Dual formulation of T_p). *Let $(\mathcal{X}, d)$ be a Polish space, $p \in [1, 2]$ and $\nu \in P_p(\mathcal{X})$. Then the following two statements are equivalent:*

(a) ν satisfies $T_p(\lambda)$;

(b) For any $\varphi \in C_b(\mathcal{X})$,

$$\begin{cases} \forall t \geq 0 \quad \displaystyle\int e^{\lambda t \, \inf_{y\in\mathcal{X}} \left[\varphi(y) + \frac{d(x,y)^p}{p}\right]} \, \nu(dx) \leq e^{\lambda\left(\frac{1}{p}-\frac{1}{2}\right) t^{\frac{2}{2-p}}} \, e^{t\lambda \int \varphi \, d\nu} \qquad (p<2); \\ \\ \displaystyle\int e^{\lambda \, \inf_{y\in\mathcal{X}} \left[\varphi(y) + \frac{d(x,y)^2}{2}\right]} \, \nu(dx) \leq e^{\lambda \int \varphi \, d\nu} \qquad (p=2). \end{cases} \tag{22.2}$$

Particular Case 22.4 (Dual formulation of T_1). Let $(\mathcal{X}, d)$ be a Polish space and $\nu \in P_1(\mathcal{X})$, then the following two statements are equivalent:

(a) ν satisfies $T_1(\lambda)$;

(b) For any $\varphi \in C_b(\mathcal{X})$,

$$\forall t \geq 0 \quad \int e^{t \, \inf_{y\in\mathcal{X}} \left[\varphi(y) + d(x,y)\right]} \, \nu(dx) \leq e^{\frac{t^2}{2\lambda}} \, e^{t \int \varphi \, d\nu}. \tag{22.3}$$

Proposition 22.5 (Tensorization of T_p). *Let $(\mathcal{X}, d)$ be a Polish space, $p \in [1,2]$ and let $\nu \in P_p(\mathcal{X})$ be a reference probability measure satisfying an inequality $T_p(\lambda)$. Then for any $N \in \mathbb{N}$, the measure $\nu^{\otimes N}$ satisfies an inequality $T_p(N^{1-\frac{2}{p}}\lambda)$ on $(\mathcal{X}^N, d_p, \nu^{\otimes N})$, where the product distance d_p is defined by*

$$d_p\big((x_1,\dots,x_N);(y_1,\dots,y_N)\big) = \Big(\sum_{i=1}^N d(x_i,y_i)^p\Big)^{\frac{1}{p}}.$$

Corollary 22.6 (T_2 inequalities tensorize exactly). *If ν satisfies $T_2(\lambda)$, then also $\mu^{\otimes N}$ satisfies $T_2(\lambda)$ on $(\mathcal{X}^N, d_2, \nu^{\otimes N})$, for any $N \in \mathbb{N}$.*

Proof of Proposition 22.2. Proposition 22.2 will be obtained as a consequence of Theorem 5.21. Recall the Legendre representation of the H-functional: For any $\lambda > 0$,

$$\begin{cases} \forall \mu \in P_2(M), \quad \displaystyle\frac{H_\nu(\mu)}{\lambda} = \sup_{\varphi \in C_b(M)} \left[\int \varphi \, d\mu - \frac{1}{\lambda} \log\left(\int_M e^{\lambda\varphi} \, d\nu \right) \right], \\ \\ \forall \varphi \in C_b(M), \quad \displaystyle\frac{1}{\lambda} \log\left(\int_M e^{\lambda\varphi} \, d\nu \right) = \sup_{\mu \in P_2(M)} \left[\int \varphi \, d\mu - \frac{H_\nu(\mu)}{\lambda} \right]. \end{cases} \tag{22.4}$$

Let us first treat the case $p = 2$. Apply Theorem 5.21 with $c(x,y) = d(x,y)^2/2$, $F(\mu) = (1/\lambda) H_\nu(\mu)$, $\Lambda(\varphi) = (1/\lambda)\log\left(\int e^{\lambda\varphi} \, d\nu\right)$. The conclusion is that ν satisfies $T_2(\lambda)$ if and only if

$$\forall \phi \in C_b(\mathcal{X}), \quad \log \int \exp\left(\lambda \int \phi \, d\nu - \lambda \phi^c \right) d\nu \leq 0,$$

i.e.

$$\int e^{-\lambda\phi^c} \, d\nu \leq e^{-\lambda \int \phi \, d\nu},$$

where $\phi^c(x) := \sup_y\big(\phi(y) - d(x,y)^2/2\big)$. Upon changing ϕ for $\varphi = -\phi$, this is the desired result. Note that the Particular Case **??** is obtained from (22.2) by choosing $p = 1$ and performing the change of variables $t \to \lambda t$.

The case $p < 2$ is similar, except that now we appeal to the equivalence between (i') and (ii') in Theorem 5.21, and choose

$$c(x,y) = \frac{d(x,y)^p}{p}; \qquad \Phi(r) = \frac{p^{\frac{2}{p}}}{2} r^{\frac{2}{p}} \, 1_{r\geq 0}; \qquad \Phi^*(t) = \left(\frac{1}{p} - \frac{1}{2}\right) t^{\frac{2}{2-p}}.$$

□

Proof of Proposition 22.5. To begin with, we need a bit of notation. Let $\mu = \mu(dx_1\, dx_2\, \dots\, dx_N)$ be a probability measure on $\mathcal{X}^N$, and let $(x_1, \dots, x_N) \in \mathcal{X}^N$ be distributed randomly according to μ. I shall write $\mu_1(dx_1)$ for the law of x_1, $\mu_2(dx_2|x_1)$ for the conditional law of x_2 given x_1, $\mu_3(dx_3|x_1, x_2)$ for the conditional law of x_3 given x_1 and x_2, etc. I shall also use the shorthand $x^i = (x_1, x_2, \dots, x_i)$, (with the convention that $x^0 = \emptyset$), and write μ^i for the law of x^i.

The proof of Proposition 22.5 is reminiscent of the strategy used to construct the Knothe–Rosenblatt coupling. First choose an optimal coupling (for the cost function $c = d^p$) between $\mu_1(dx_1)$ to $\nu(dy_1)$, call it $\pi_1(dx_1\, dy_1)$. Then for each x_1, choose an optimal between $\mu_2(dx_2|x_1)$ and $\nu(dy_2)$, call it $\pi_2(dx_2\, dy_2|x_1)$. Then for each (x_1, x_2), choose an optimal coupling between $\mu_3(dx_3|x_1, x_2)$ and $\nu(dy_3)$, call it $\pi_3(dx_3\, dy_3|x_1, x_2)$; etc. In the end, glue these plans together to get a coupling

$$\pi(dx_1\, dy_1\, dx_2\, dy_2 \dots\, dx_N\, dy_N) = \pi_1(dx_1\, dy_1)\, \pi_2(dx_2\, dy_2|x_1)\, \pi_3(dx_3\, dy_3|x_1, x_2) \dots \\ \dots\, \pi_N(dx_N\, dy_N|x_1, \dots, x_{N-1}).$$

In more compact notation,

$$\pi(dx\, dy) = \pi_1(dx_1\, dy_1)\, \pi_2(dx_2\, dy_2|x^1) \dots\, \pi_N(dx_N\, dy_N|x^{N-1}).$$

Here something should be said about the measurability, since there is a priori no canonical way to choose $\pi_i(\,\cdot\,|x^{i-1})$ as a measurable function of x^{i-1}. But it is a general fact that if $x \to \mu^x$ is a measure-valued measurable map, and ν is another measure, then for each x one can choose an optimal transference plan $\pi = \pi^x$ between μ^x and ν, in a measurable way. To see this, let Π_ν be the set of all optimal transference plans whose second marginal is ν; and let f be the map which to $\pi \in \Pi_\nu$ associates its first marginal. Obviously f is continuous; by Theorem 4.1 it is surjective; and by Corollary 5.19 all preimages $f^{-1}(\mu)$ are compact. So the conclusion follows from the measurable selection theorem.

By the definition of d_p,

$$\begin{aligned} \mathbb{E}_\pi\, d_p(x, y)^p &= \sum_{i=1}^N \mathbb{E}_\pi\, d(x_i, y_i)^p \\ &= \sum_{i=1}^N \int \Big[\mathbb{E}_{\pi(\,\cdot\,|x^{i-1})}\, d(x_i, y_i)^p\Big]\, \pi^{i-1}(dx^{i-1}\, dy^{i-1}) \\ &= \sum_{i=1}^N \int \Big[\mathbb{E}_{\pi(\,\cdot\,|x^{i-1})}\, d(x_i, y_i)^p\Big]\, \mu^{i-1}(dx^{i-1}), \end{aligned} \tag{22.5}$$

where of course

$$\pi^i(dx^i\, dy^i) = \pi_1(dx_1\, dy_1)\, \pi_2(dx_2\, dy_2|x^1) \dots\, \pi_i(dx_i\, dy_i|x^{i-1}).$$

For each i and each $x^{i-1} = (x_1, \dots, x_{i-1})$, the measure $\pi(\,\cdot\,|x^{i-1})$ is an optimal transference plan between its marginals. So the right-hand side of (22.5) can be rewritten as

$$\sum_{i=1}^N \int W_p\big(\mu_i(\,\cdot\,|x^{i-1}), \nu\big)^p\, \mu^{i-1}(dx^{i-1}).$$

Since this cost is achieved for the transference plan π, we otain the key estimate

$$W_p(\mu, \nu^{\otimes N})^{\otimes p} \le \sum_{i=1}^{N} \int W_p\big(\mu_i(\,\cdot\,|x^{i-1}), \nu\big)^p \, \mu^{i-1}(dx^{i-1}). \tag{22.6}$$

By assumption, ν satisfies $T_p(\lambda)$, so the right-hand side in (22.6) can be bounded above by

$$\sum_i \int \left(\frac{2}{\lambda} H_\nu\big(\mu_i(\,\cdot\,|x^{i-1})\big) \right)^{\frac{p}{2}} \mu^{i-1}(dx^{i-1}). \tag{22.7}$$

Since $p \le 2$, we can apply Hölder's inequality, in the form $\sum_{i \le N} a_i^{p/2} \le N^{1-p/2} (\sum a_i)^{p/2}$, and bound (22.6) by

$$N^{1-\frac{p}{2}} \left(\frac{2}{\lambda} \right)^{\frac{p}{2}} \int \left(\sum_{i=1}^{N} H_\nu\big(\mu_i(\,\cdot\,|x^{i-1})\big) \right) \mu^{i-1}(dx^{i-1}). \tag{22.8}$$

But the formula of **additivity of entropy**, proven as a lemma below, states that

$$\sum_{1 \le i \le N} \int H_\nu\big(\mu_i(dx_i|x^{i-1})\big) \, \mu^{i-1}(dx^{i-1}) = H_{\nu^{\otimes N}}(\mu). \tag{22.9}$$

Putting back all the previous bounds together, we end up with

$$W_p(\mu, \nu)^p \le N^{1-\frac{p}{2}} \left(\frac{2}{\lambda} \right)^{\frac{p}{2}} H_\nu(\mu)^{\frac{p}{2}},$$

which is the same as the desired inequality. □

Remark 22.7. The same proof shows that the inequality

$$\forall \mu \in P(\mathcal{X}), \qquad C(\mu, \nu) \le H_\nu(\mu)$$

implies

$$\forall \mu \in P(\mathcal{X}^N), \qquad C^N(\mu, \nu) \le H_{\nu^{\otimes N}}(\mu),$$

where C^N is the optimal transport cost associated with the cost function

$$c^N(x, y) = \sum c(x_i, y_i)$$

on $\mathcal{X}^N$.

The following important lemma was used in the course of the proof of Proposition 22.5.

Lemma 22.8 (additivity of the entropy). *Let $\mathcal{X}$ be a Polish space, $\nu \in P(\mathcal{X})$, $N \in \mathbb{N}$, and $\mu \in P(\mathcal{X}^N)$. Then, with the same notation as in the beginning of the proof of Proposition 22.5,*

$$H_{\nu^{\otimes N}}(\mu) = \sum_{1 \le i \le N} \int H_\nu\big(\mu_i(dx_i|x^{i-1})\big) \, \mu^{i-1}(dx^{i-1}). \tag{22.10}$$

Proof of Lemma 22.8. By induction, it suffices to treat the case $N = 2$. Let $\rho = \rho(x_1, x_2)$ be the density of μ with respect to $\nu \otimes \nu$. By an easy approximation argument based on the monotone convergence theorem, it is sufficient to establish (22.10) in the case when ρ is bounded.

The conditional measure $\mu_2(dx_2|x_1)$ has density $\rho(x_1,x_2)/(\int \rho(x_1,x_2)\,\nu(dx_2))$, and the measure $\mu_1(dx_1)$ has density $\int \rho(x_1,x_2)\,\nu(dx_2)$. From this and the additive properties of the logarithm, we deduce

$$\begin{aligned}
&\int H_\nu\big(\mu_2(\,\cdot\,|x_1)\big)\,\mu_1(dx_1)\\
&= \int \left(\int \frac{\rho(x_1,x_2)}{\int \rho(x_1,x_2')\,\nu(dx_2')} \log \frac{\rho(x_1,x_2)}{\int \rho(x_1,x_2')\,\nu(dx_2')}\,\nu(dx_2)\right)\left(\int \rho(x_1,x_2')\,\nu(dx_2')\right)\nu(dx_1)\\
&= \iint \rho(x_1,x_2)\log\rho(x_1,x_2)\,\nu(dx_2)\,\nu(dx_1) - \int \Big(\int \rho(x_1,x_2)\,\nu(dx_2)\Big)\log\Big(\int \rho(x_1,x_2)\,\nu(dx_2)\Big)\,\nu(dx_1)\\
&= H_\nu(\mu) - H_\nu(\mu_1).
\end{aligned}$$

This concludes the proof. □

Gaussian concentration

Gaussian concentration is a loose terminology meaning that some reference measure enjoys properties of concentration of measure which are similar to those of the Gaussian measure. In this section we shall see that a certain form of Gaussian concentration is *equivalent* to a T_1 inequality.

Theorem 22.9 (Gaussian concentration). *Let $(\mathcal{X},d)$ be a Polish space, equipped with a reference probability measure ν. Then the following properties are equivalent:*

(i) ν lies in $P_1(\mathcal{X})$ and satisfies a T_1 inequality;

(ii) There is $\lambda>0$ such that for any $\varphi\in C_b(\mathcal{X})$,

$$\forall t\ge 0 \quad \int e^{t\,\inf_{y\in\mathcal{X}}\left[\varphi(y)+d(x,y)\right]}\,\nu(dx) \le e^{\frac{t^2}{2\lambda}}\,e^{t\int\varphi\,d\nu}.$$

(iii) There is a constant $C>0$ such that for any Borel subset A of $\mathcal{X}$,

$$\nu[A]\ge\frac12 \Longrightarrow \quad \nu[A^r]\ge 1-e^{-C\,r^2};$$

(iv) There is a constant $C>0$ such that

$$\forall f\in L^1(\nu)\cap \mathrm{Lip}(\mathcal{X}),\ \forall\varepsilon>0,\qquad \nu\Big[\big\{x\in\mathcal{X};\ f(x)\ge\int f\,d\nu+\varepsilon\big\}\Big]\le\exp\left(-C\frac{\varepsilon^2}{\|f\|^2_{\mathrm{Lip}}}\right);$$

(v) There is a constant $C>0$ such that

$$\forall f\in L^1(\nu)\cap\mathrm{Lip}(\mathcal{X}),\ \forall\varepsilon>0,\ \forall N\in\mathbb{N},\quad \nu^{\otimes N}\Big[\big\{x\in\mathcal{X}^N;\ \frac1N\sum_{i=1}^N f(x_i)\ge\int f\,d\nu+\varepsilon\big\}\Big]\le\exp\left(-C\frac{N\,\varepsilon^2}{\|f\|^2_{\mathrm{Lip}}}\right);$$

(vi) There is a constant $C>0$ such that

$$\forall f\in\mathrm{Lip}(\mathcal{X}),\ \forall\varepsilon>0,\quad \nu\Big[\big\{x\in\mathcal{X};\ f(x)\ge m_f+\varepsilon\big\}\Big]\le\exp\left(-C\frac{\varepsilon^2}{\|f\|^2_{\mathrm{Lip}}}\right),$$

where m_f stands for any median of f;

(vii) For any $x_0 \in \mathcal{X}$ there is a constant $a > 0$ such that

$$\int e^{a\, d(x_0,x)^2}\, \nu(dx) < +\infty;$$

(viii) There exists $a > 0$ such that

$$\int e^{a\, d(x,y)^2}\, \nu(dx)\, \nu(dy) < +\infty;$$

(ix) There exist $x_0 \in \mathcal{X}$ and $a > 0$ such that

$$\int e^{a\, d(x_0,x)^2}\, \nu(dx) < +\infty.$$

Proof of Theorem 22.9. We shall prove (i) ⇒ (ii) ⇒ (iv) ⇒ (vii), (i) ⇒ (v) ⇒ (iv), (i) ⇒ (iii) ⇒ (vi) ⇒ (vii), (vii) ⇒ (viii) ⇒ (ix) ⇒ (i), and this will establish the theorem.

(i) ⇒ (ii) was already seen in Particular Case 22.4.

To prove (ii) ⇒ (iv), first note that it suffices to treat the case $\|f\|_{\mathrm{Lip}} = 1$ (replace ε by $\varepsilon/\|f\|_{\mathrm{Lip}}$ and f by $f/\|f\|_{\mathrm{Lip}}$). Then if f is 1-Lipschitz,

$$\inf_{y\in\mathcal{X}} \big[f(y) + d(x,y)\big] = f(x),$$

so (ii) implies

$$\int e^{t\, f(x)}\, \nu(dx) \le e^{\frac{t^2}{2\lambda}}\, e^{t \int f\, d\nu}.$$

With the shorthand $\langle f \rangle = \int f\, d\nu$, this is the same as

$$\int e^{t\,(f - \langle f\rangle)}\, d\nu \le e^{\frac{t^2}{2\lambda}}.$$

Then by the exponential Chebyshev inequality,

$$\nu\Big[\big\{f - \langle f\rangle \ge \varepsilon\big\}\Big] \le e^{-t\varepsilon} e^{\frac{t^2}{2\lambda}};$$

and (v) follows by taking the infimum over $t > 0$. Note that $C = \lambda/2$ does the job.

Now let us prove (iv) ⇒ (vii). Let ν satisfy (iv). First we shall check that $d(\cdot, x_0) \in L^1(\nu)$. Let $m \in \mathbb{N}$, and let $f_m = d(\cdot, x_0) \wedge m$; then $f_m \in L^1(\nu) \cap \mathrm{Lip}(\mathcal{X})$, so

$$\nu\big[f_m \ge s + \int f_m\, d\nu\big] \le e^{-C\, s^2}.$$

It follows that for any $A \le m$,

$$\begin{aligned}
\int f_m^2\, d\nu &= \int_0^{+\infty} 2s\, \nu[f_m \ge s]\, ds \\
&\le \int_0^A 2s\, \nu[f_m \ge s]\, ds + \int_A^{\int f_m\, d\nu} 2s\, \nu[f_m \ge s]\, ds + \int_{\int f_m\, d\nu}^{+\infty} 2s\, \nu\big[f_m \ge s\big]\, ds \\
&\le 2A + \nu[f_m \ge A] \int_A^{\int f_m\, d\nu} 2s\, ds + \int_0^{+\infty} 2\big(s + \int f_m\, d\nu\big)\, \nu\Big[f_m \ge s + \int f_m\, d\nu\Big]\, ds \\
&\le 2A + \nu[f_m \ge A] \left(\int f_m\, d\nu\right)^2 + \int_0^{+\infty} 2s\, e^{-C\, s^2}\, ds + 2\left(\int f_m\, d\nu\right) \int_0^{+\infty} e^{-C\, s^2}\, ds \\
&\le 2A + \nu[f \ge A] \left(\int f_m\, d\nu\right)^2 + \int_0^{+\infty} 2s\, e^{-C\, s^2}\, ds + \frac{1}{4}\left(\int f_m\, d\nu\right)^2 + 8\left(\int_0^{+\infty} e^{-C\, s^2}\, ds\right)^2 \\
&\le 2A + \left(\int f_m^2\, d\nu\right) \left(\nu[f \ge A] + \frac{1}{4}\right) + \overline{C},
\end{aligned}$$

where $\overline{C} = \int_0^{+\infty} 2s\, e^{-C\,s^2}\, ds + 8\left(\int_0^{+\infty} e^{-C\,s^2}\, ds\right)^2$ is a finite constant. If A is large enough, then $\nu[f \geq A] \leq 1/4$, and then the above inequality implies $\int f_m^2\, d\nu \leq 2(2A + \overline{C})$. By taking $m \to \infty$ we deduce that $\int f^2\, d\nu < +\infty$, in particular $f \in L^1(\nu)$. So we can apply directly (iv) to $f = d(\cdot, x_0)$, and it follows that for any $a < C$,

$$\begin{aligned}\int e^{a\,d(x,x_0)^2}\,\nu(dx) &= \int_0^{+\infty} 2ase^{as^2}\,\nu[f \geq s]\,ds \\ &= \int_0^{\int f\,d\nu} 2ase^{as^2}\,\nu[f \geq s]\,ds + \int_0^{+\infty} 2a\Big(s + \int f\,d\nu\Big)\,e^{a\left(s+\int f\,d\nu\right)^2}\,\nu\Big[f \geq s + \int f\,d\nu\Big]\,ds \\ &\leq 2a\left(\int f\,d\nu\right)e^{a\left(\int f\,d\nu\right)^2} + \int_0^{+\infty} 2a\Big(s + \int f\,d\nu\Big)\,e^{a\left(s+\int f\,d\nu\right)^2} e^{-\,Cs^2}\,ds \;\; < +\infty.\end{aligned}$$

This proves (vii).

The next implication is (i) $\Rightarrow$ (v). If ν satisfies $T_1(\lambda)$, then by Proposition 22.5 $\nu^{\otimes N}$ satisfies $T_1(\lambda/N)$ on $\mathcal{X}^N$ equipped with the distance $d_1(x,y) = \sum d(x_i, y_i)$. Let $F : \mathcal{X}^N \to \mathbb{R}$ be defined by

$$F(x) = \frac{1}{N}\sum_{i=1}^N f(x_i).$$

If f is Lipschitz then $\|F\|_{\mathrm{Lip}} = \|f\|_{\mathrm{Lip}}/N$. Moreover, $\int F\,d\nu^{\otimes N} = \int f\,d\nu$. So if we apply (iv) with $\mathcal{X}$ replaced by $\mathcal{X}^N$ and f replaced by F, we obtain

$$\begin{aligned}\nu^{\otimes N}\Big[\Big\{x \in \mathcal{X}^N;\ \frac{1}{N}\sum_{i=1}^N f(x_i) \geq \int f\,d\nu + \varepsilon\Big\}\Big] &= \nu^{\otimes N}\Big[\Big\{x \in \mathcal{X}^N;\ F(x) \geq \int F\,d\nu + \varepsilon\Big\}\Big] \\ &\leq \exp\left(-\,(C/N)\frac{\varepsilon^2}{(\|f\|_{\mathrm{Lip}}/N)^2}\right) \\ &= \exp\left(-\,C\frac{N\varepsilon^2}{\|f\|_{\mathrm{Lip}}^2}\right),\end{aligned}$$

where $C = \lambda/2$ (Cf. the remark at the end of the proof of (i) $\Rightarrow$ (iv)).

The implication (v) $\Rightarrow$ (iv) is trivial.

Let us now consider the implication (i) $\Rightarrow$ (iii). Assume that

$$\forall \mu \in P_1(\mathcal{X}), \qquad W_1(\mu,\nu) \leq C\sqrt{H_\nu(\mu)}. \tag{22.11}$$

Choose A with $\nu[A] \geq 1/2$, and $\mu = (1_A\,\nu)/\nu[A]$, $\widetilde{\mu} = (1_{\mathcal{X}\setminus A^r}\,\nu)/\nu[\mathcal{X} \setminus A^r]$. It is an immediate computation that

$$H_\nu(\mu) = \log\frac{1}{\nu[A]} \leq \log 2, \qquad H_\nu(\widetilde{\mu}) = \log\left(\frac{1}{1-\nu[A^r]}\right).$$

By (22.11) and the triangular inequality for the distance W_1,

$$W_1(\mu,\widetilde{\mu}) \leq W_1(\mu,\nu) + W_1(\widetilde{\mu},\nu) \leq C\sqrt{\log 2} + C\sqrt{\log\left(\frac{1}{1-\nu[A^r]}\right)}. \tag{22.12}$$

On the other hand, it is obvious that $W_1(\mu,\widetilde{\mu}) \geq r$ (all the mass has to go from A to $\mathcal{X} \setminus A^r$, so each unit of mass should travel a distance at least r). So (22.12) implies

$$r \le C\sqrt{\log 2} + C\sqrt{\log\left(\frac{1}{1-\nu[A^r]}\right)},$$

from which it is easy to deduce

$$\nu[A^r] \ge 1 - \exp\left[-\left(\frac{r}{C} - \sqrt{\log 2}\right)^2\right].$$

This establishes a bound of the type $\nu[A^r] \ge 1 - a e^{-C\,r^2}$. Property (iii) follows. (To get rid of the constant a, note that $\nu[A^r] \ge \min(1/2, 1 - a e^{-c r^2}) \ge 1 - e^{-c' r^2}$ for c' well-chosen).

To prove (iii) $\Rightarrow$ (vi), let $A = \{y;\, f(y) \le m_f\}$. By definition of a median, A has probability at least $1/2$. So if $f(x) \ge m_f + \varepsilon$, it follows that $f(x) - f(y) \ge \varepsilon$ for any $y \in A$, so $d(x,y) \ge \varepsilon / \|f\|_{\mathrm{Lip}}$. In other words, $d(x,A) \ge \varepsilon/\|f\|_{\mathrm{Lip}}$ and $\{f \ge m_f + \varepsilon\}$ is included in $\mathcal{X} \setminus A^r$, $r = \varepsilon/\|f\|_{\mathrm{Lip}}$. Then (iii) leads to (vi).

To show (vi) $\Rightarrow$ (vii), let A be a compact set such that $\nu[A] \ge 1/2$; let also $x_0 \in A$, and let R be the diameter of A. Let further $f(x) = d(x,A)$; then f is a 1-Lipschitz function admitting 0 for median. So (vi) implies

$$\nu\big[d(x,x_0) \ge R + r\big] \le \nu[d(x,A) \ge r] \le e^{-C\,r^2}.$$

It follows that for any $a < C$,

$$\begin{aligned}\int e^{a\,d(x,x_0)^2}\,\nu(dx) &= \int_0^{+\infty} \nu\big[d(x,x_0)^2 \ge s\big]\, 2ase^{as^2}\,ds \\ &\le \int_0^R 2ase^{as^2}\,ds + \int_R^{+\infty} \nu\big[d(x,x_0)^2 \ge s\big]\, 2ase^{as^2}\,ds \\ &\le Re^{aR^2} + \int_R^{\infty} e^{-C(s-R)^2} 2ase^{as^2}\,ds \;\; < +\infty.\end{aligned}$$

To prove (vii) $\Rightarrow$ (viii), pick up any $x_0 \in \mathcal{X}$ and write

$$\int e^{a\,d(x,y)^2}\,\nu(dx)\,\nu(dy) \le \int e^{2a\,d(x,x_0)^2 + 2a\,d(x_0,y)^2}\,\nu(dx)\,\nu(dy) = \left(\int e^{2a\,d(x,x_0)^2}\,\nu(dx)\right)^2.$$

The implication (viii) $\Rightarrow$ (ix) is obvious.

It only remains to establish (ix) $\Rightarrow$ (i). If ν satisfies (ix), then obviously $\nu \in P_1(\mathcal{X})$. To prove that ν satisfies T_1, we shall establish the **weighted Csiszár–Kullback–Pinsker inequality**

$$\big\|d(x_0,\cdot)\,(\mu-\nu)\big\|_{TV} \;\le\; \sqrt{2}\left(1 + \log\int_{\mathcal{X}} e^{d(x_0,x)^2}\,d\nu(x)\right)^{1/2} \sqrt{H_\nu(\mu)}. \tag{22.13}$$

Inequality (22.13) implies the T_1 inequality, since Theorem 6.12 gives

$$W_1(\mu,\nu) \le \big\|d(x_0,\cdot)\,(\mu-\nu)\big\|_{TV}.$$

So we turn to the proof of (22.13). We may assume that μ is absolutely continuous with respect to ν, otherwise (22.13) is trivial. Let then f be the density of μ, and let $u = f - 1$, so that

$$\mu \;=\; (1+u)\,\nu;$$

note that $u \geq -1$ and $\int u \, d\nu = 0$. We also define

$$h(v) := (1+v)\log(1+v) - v \geq 0, \qquad v \in [-1, +\infty);$$

so that

$$H_\nu(\mu) = \int_{\mathcal{X}} h(u) \, d\nu. \tag{22.14}$$

Finally, let

$$\varphi(x) = d(x_0, x).$$

Since $h(0) = h'(0) = 0$, Taylor's formula (with integral remainder) yields

$$h(u) = u^2 \int_0^1 \frac{1-t}{1+tu} \, dt,$$

so

$$H_\nu(\mu) = \int_{\mathcal{X}} \int_0^1 \frac{u^2(x)\,(1-t)}{1+tu(x)} \, d\nu(x) \, dt.$$

On the other hand, by Cauchy–Schwarz inequality on $(0,1) \times \mathcal{X}$

$$\begin{aligned}\left(\int_0^1 (1-t)\, dt\right)^2 \left(\int_{\mathcal{X}} \varphi |u| \, d\nu\right)^2 &= \left(\int_{(0,1)\times\mathcal{X}} (1-t)\varphi|u| \, d\nu \, dt\right)^2 \\ &\leq \left(\iint (1-t)\,(1+tu)\,\varphi^2 \, d\nu \, dt\right) \left(\iint \frac{1-t}{1+tu} |u|^2 \, d\nu \, dt\right);\end{aligned}$$

thus

$$\left(\int \varphi \, |u| \, d\nu\right)^2 \leq C H_\nu(\mu),$$

where

$$C := \frac{\displaystyle\iint (1-t)\,(1+tu)\,\varphi^2 \, d\nu \, dt}{\displaystyle\left(\int_0^1 (1-t)\, dt\right)^2}. \tag{22.15}$$

We decompose the numerator as follows:

$$\begin{aligned}\iint (1-t)\,(1+tu)\,\varphi^2 \, d\nu \, dt &= \int (1-t) t \, dt \int (1+u)\,\varphi^2 \, d\nu + \int (1-t)^2 \, dt \int \varphi^2 \, d\nu \\ &= \frac{1}{6} \int \varphi^2 \, d\mu + \frac{1}{3} \int \varphi^2 \, d\nu. \end{aligned} \tag{22.16}$$

From the Legendre representation of the H functional,

$$\int \varphi^2 \, d\mu \leq H_\nu(\mu) + \log \int e^{\varphi^2} \, d\nu, \tag{22.17}$$

and Jensen's inequality, in the form

$$\int \varphi^2 \, d\nu \leq \log \int e^{\varphi^2} \, d\nu, \tag{22.18}$$

we deduce that the right-hand side of (22.16) is bounded above by

$$\frac{1}{6}H_\nu(\mu) + \frac{1}{2}\log \int e^{\varphi^2}\, d\nu.$$

Plugging this into (22.15), we conclude that

$$\left(\int \varphi\, |u|\, d\nu\right)^2 \le \left(\frac{2}{3}H + 2L\right) H, \tag{22.19}$$

where H stands for $H_\nu(\mu)$ and L for $\log \int e^{\varphi^2}\, d\nu$.

The preceding bound is relevant only for "small" values of H. To handle large values, we write

$$\begin{aligned}\left(\int \varphi|u|\, d\nu\right)^2 &\le \int \varphi^2|u|\, d\nu \int |u|\, d\nu \\ &\le \left(\int \varphi^2\, d\mu + \int \varphi^2\, d\nu\right)\left(\int d\mu + \int d\nu\right) \\ &\le (H + 2L)\, 2\end{aligned}$$

where we have successively used Cauchy–Schwarz inequality, the inequality $|u| \le 1 + u + 1$ on $[-1, +\infty)$ (which results in $|u|\,\nu \le \mu + \nu$), and finally (22.17) and (22.18).

Combining this with (22.19), we obtain

$$\left(\int \varphi|u|\, d\nu\right)^2 \le \min\left((2H)\big(\frac{H}{3} + L\big)\,,\, 2(H + 2L)\right).$$

From the elementary inequality

$$\min\left(at^2 + bt,\, t + d\right) \le Mt, \qquad M = \frac{1}{2}\left\{1 + b + \sqrt{(b-1)^2 + 4ad}\right\}$$

we get

$$\int \varphi|u|\, d\nu \le m\sqrt{H(\mu|\nu)}$$

where

$$m \le \sqrt{1 + L + \sqrt{(L-1)^2 + \frac{8}{3}L}} \le \sqrt{2}\,\sqrt{L+1}.$$

This concludes the proof. □

Talagrand inequalities from Ricci curvature bounds

In the previous section we have focused on T_1 inequalities; now we consider T_2 inequalities (Talagrand inequalities). The most simple criterion for T_2 to hold is in terms of *Ricci curvature bounds*:

Theorem 22.10 ($\mathrm{CD}(K,\infty)$ **implies** $T_2(K)$**).** *Let M be a Riemannian manifold, equipped with a reference probability measure $\nu = e^{-V}\mathrm{vol}$, satisfying the curvature-dimension bound $\mathrm{CD}(K,\infty)$. Then ν lies in $P_2(M)$ and satisfies the Talagrand inequality $T_2(K)$. In particular, ν satisfies Gaussian concentration bounds.*

Proof of Theorem 22.10. It follows from Corollary 19.13 that ν lies in $P_2(M)$; then the inequality $T_2(K)$ follows from Theorem 20.7 with $\mu_0 = \nu$ and $\mu_1 = \mu$. Since $T_2(K)$ implies $T_1(K)$, Theorem 22.9 shows that ν satisfies Gaussian concentration bounds. □

Example 22.11. The standard Gaussian γ on $\mathbb{R}^N$ satisfies $\mathrm{CD}(1, \infty)$, and therefore $T_2(1)$ too. This is independent of N.

Remark 22.12. If ν satisfies $T_2(\lambda)$, then also $\nu^{\otimes N}$ satisfies $T_2(\lambda)$, *independently of* N; so one might hope to improve the concentration inequality appearing in Theorem 22.9(v). But now the space $\mathcal{X}^N$ should be equipped with the d_2 distance, for which the function $F : x \to (1/N) \sum \varphi(x_i)$ is only $\sqrt{N}$-Lipschitz! In the end, the concentration inequality derived from T_2 is similar to the one derived from T_1. This is not in contradiction with the fact that T_2 is quite stronger than T_1; it just shows that we do not see the difference when we consider observables of the particular form $(1/N) \sum \varphi(x_i)$.

Relation with log Sobolev and Poincaré inequalities

So far we learnt that logarithmic Sobolev inequalities follow from curvature bounds, and that Talagrand inequalities also follow from the same bounds. We also learnt from Chapter 21 that logarithmic Sobolev inequalities imply Poincaré inequalities. Actually, Talagrand inequalities are *intermediate* between these two inequalities: a logarithmic Sobolev inequality implies a Talagrand inequality, which in turn implies a Poincaré inequality. In some sense however, Talagrand is closer to logarithmic Sobolev than to Poincaré: For instance, in nonnegative curvature, the validity of the Talagrand inequality is equivalent to the validity of the logarithmic Sobolev inequality up to a degradation of the constant by a factor 1/4.

To establish these properties, we shall use, for the first time in this course, a **semigroup argument**. As understood by Ledoux, it is convenient to consider inequality (22.2) from a dynamical point of view, with the help of the (forward) Hamilton–Jacobi semigroup defined as in Chapter 7 by

$$\begin{cases} H_0\, \varphi = \varphi, \\ \\ (H_t\, \varphi)(x) = \inf_{y \in M} \left[\varphi(x) + \dfrac{d(x,y)^2}{2t} \right] \qquad (t > 0,\ x \in M). \end{cases} \tag{22.20}$$

The next proposition, stated here without proof, summarizes some of the nice properties of the semigroup $(H_t)_{t \geq 0}$. Recall the notation $|\nabla^- f|$ from (20.2).

Proposition 22.13 (Properties of the quadratic Hamilton–Jacobi semigroup). *Let f be a bounded continuous function on a Riemannian manifold M. Then*

(i) For all $s, t \geq 0$, $H_t H_s f = H_{t+s} f$.

(ii) For all $x \in M$, $\inf f \leq (H_t f)(x) \leq f(x)$; moreover, the infimum in (22.49) *can be restricted to $y \in B(x, \sqrt{Ct})$, where $C := 2(\sup f - \inf f)$.*

(iii) For all $t > 0$, $H_t f$ is locally Lipschitz on M.

(iv) For all $x \in M$, $(H_t f)(x)$ is a nonincreasing function of t, that converges monotonically to $f(x)$ as $t \to 0$. In particular, $\lim_{t \to 0} H_t f = f$, locally uniformly in x.

(v) For all $t \geq 0$, $s > 0$, $x \in M$,

$$\frac{|H_{t+s} f(x) - H_t f(x)|}{s} \leq \frac{\|H_t f\|^2_{\mathrm{Lip}(B(x, \sqrt{Cs}))}}{2}.$$

(vi) For all $x \in M$ and $t \geq 0$,

$$\liminf_{s \to 0^+} \frac{(H_{t+s} f)(x) - (H_t f)(x)}{s} \geq - \frac{|\nabla^- H_t f|^2(x)}{2}. \tag{22.21}$$

(vii) For all $x \in M$ and $t > 0$,

$$\lim_{s \to 0^+} \frac{(H_{t+s} f)(x) - (H_t f)(x)}{s} = - \frac{|\nabla^- H_t f|^2(x)}{2}. \tag{22.22}$$

Remark 22.14. The last part of the theorem shows that H_t has a built-in regularization effect which goes further than Lipschitz regularity; property (vii) is not *a priori* true for $t = 0$, even if f is Lipschitz. In some sense, for any $t > 0$, $H_t f$ is "semi-concave". Statement (vii) is the most tricky part of Proposition 22.13, and actually the only one in which some smoothness of M is used.

Now we are ready for the main result of this section.

Theorem 22.15 (Logarithmic Sobolev implies T_2 implies Poincaré). *Let M be a Riemannian manifold equipped with a reference measure $\nu \in P_2(M)$. Then*

(i) If ν satisfies a logarithmic Sobolev inequality with constant $K > 0$, then it also satisfies a Talagrand inequality with constant K.

(ii) If ν satisfies a Talagrand inequality with constant $K > 0$, then it also satisfies a Poincaré inequality with constant K.

Remark 22.16. Theorem 22.15 has the important advantage over Theorem 22.10 that logarithmic Sobolev inequalities are somewhat easy to perturb (recall Remark 21.5), while there is no good perturbation criterion for T_2. Essentially, the best known partial result in that direction is as follows: if ν satisfies T_2 and $\widetilde{\nu} = e^{-v} \nu$ with v bounded, then there is a constant C such that

$$\forall \mu \in P_2(M), \quad W_2(\mu, \nu) \leq C \Big(\sqrt{H_\nu(\mu)} + H_\nu(\mu)^{\frac{1}{4}} \Big). \tag{22.23}$$

Remark 22.17. Part (ii) of Theorem 22.15 shows that the T_2 inequality on a Riemannian manifold contains spectral information, and imposes many restrictions on the shape of measures satisfying T_2. For instance, it is impossible for the support of such a measure to have two disjoint components. (Take $u = a$ one one component, $u = b$ on another, $u = 0$ elsehwere, where a and b are two constants chosen in such a way that $\int u \, d\nu = 0$. Then $\int |\nabla u|^2 \, d\nu = 0$ while $\int u^2 \, d\nu > 0$.) This remark shows that T_2 does not result from just decay estimates, in contrast with T_1.

Proof of Theorem 22.15, part (i). Let ν satisfy a logarithmic Sobolev inequality with constant $K > 0$. By the dual reformulation of $T_2(K)$ (Proposition 22.2 for $p = 2$), it is sufficient to show that, for any $g \in C_b(\mathcal{X})$,

$$\int_M e^{K(Hg)} \, d\nu \leq e^{\int_M g \, d\nu}, \tag{22.24}$$

where

$$(Hg)(x) = \inf_{y \in M} \Big[g(y) + \frac{d(x,y)^2}{2} \Big].$$

Define

$$\phi(t) = \frac{1}{Kt} \log \left(\int_M e^{KtH_t g} \, d\nu \right). \tag{22.25}$$

Since g is bounded, Proposition 22.13(ii) implies that $H_t g$ is bounded, uniformly in t. Thus

$$e^{KtH_t g} = 1 + Kt \int_M H_t g \, d\nu + O(t^2) \tag{22.26}$$

and

$$\phi(t) = \int_M H_t g \, d\nu + O(t). \tag{22.27}$$

By Proposition 22.13(iv), $H_t g$ converges pointwise to g as $t \to 0^+$; so by the dominated convergence theorem,

$$\lim_{t \to 0^+} \phi(t) \; = \; \int_M g \, d\nu. \tag{22.28}$$

So it all amounts to show that $\phi(1) \leq \lim_{t \to 0^+} \phi(t)$, and this will obviously be true if $\phi(t)$ is nonincreasing in t. To prove this, we shall compute the time-derivative $\phi'(t)$. We shall go slowly, so the hasty reader may go directly to the result, which is formula (22.39) below.

Let $t \in (0, 1]$ be given. For $s > 0$, we have

$$\begin{aligned} \frac{\phi(t+s) - \phi(t)}{s} \; &= \frac{1}{s} \left(\frac{1}{K(t+s)} - \frac{1}{Kt} \right) \log \int_M e^{K(t+s)H_{t+s} g} \, d\nu \\ &+ \; \frac{1}{Kts} \left(\log \int_M e^{K(t+s)H_{t+s} g} \, d\nu - \log \int_M e^{KtH_t g} \, d\nu \right). \end{aligned} \tag{22.29}$$

As $s \to 0^+$, $e^{K(t+s)H_{t+s} g}$ converges pointwise to $e^{Kt\, H_t g}$, and is uniformly bounded. So the first term in the right-hand side of (22.29) converges, as $s \to 0^+$, to

$$- \frac{1}{Kt^2} \; \log \left(\int_M e^{Kt\, H_t g} \, d\nu \right). \tag{22.30}$$

On the other hand, the second term in the right-hand side of (22.29) converges to

$$\frac{1}{Kt \int e^{Kt\, H_t g} \, d\nu} \; \lim_{s \to 0^+} \; \left[\frac{1}{s} \left(\int_M e^{K(t+s)H_{t+s} g} \, d\nu - \int_M e^{Kt\, H_t g} \, d\nu \right) \right], \tag{22.31}$$

provided that the latter limit exists.

To evaluate the limit in (22.31), we rewrite the expression inside the square brackets as

$$\int_M \left(\frac{e^{K(t+s)H_{t+s} g} - e^{Kt\, H_{t+s} g}}{s} \right) d\nu + \int_M \left(\frac{e^{Kt\, H_{t+s} g} - e^{Kt\, H_t g}}{s} \right) d\nu. \tag{22.32}$$

The integrand of the first term in (22.32) can be rewritten as $(e^{Kt\, H_{t+s} g})(e^{Ks\, H_{t+s} g} - 1)/s$, which is uniformly bounded and converges pointwise to $(e^{K\, H_t g}) Kt\, H_t g$ as $s \to 0^+$. So the first integral in (22.32) converges to $\int_M (K\, H_t g) e^{Kt\, H_t g} \, d\nu$.

Now we consider the second term of (22.32). By Proposition 22.13(vii), for each $x \in M$,

$$H_{t+s} g(x) = H_t g(x) - s \left(\frac{|\nabla^- H_t g(x)|^2}{2} + o(1) \right), \tag{22.33}$$

and therefore

$$\lim_{s\to 0^+} \frac{e^{KtH_{t+s}g(x)} - e^{Kt\,H_tg(x)}}{s} \;=\; -\,Kte^{Kt\,H_tg}\,\frac{|\nabla^- H_tg(x)|^2}{2}. \tag{22.34}$$

On the other hand, parts (iv) and (v) of Proposition 22.13 imply that

$$H_{t+s}g = H_tg + O(s). \tag{22.35}$$

Since $H_tg(x)$ is uniformly bounded in t and x,

$$\frac{e^{KtH_{t+s}g} - e^{Kt\,H_tg}}{s} = O(1) \qquad \text{as } s \to 0^+. \tag{22.36}$$

By (22.34), (22.36) and the dominated convergence theorem,

$$\lim_{s\to 0^+} \int_M \left(\frac{e^{KtH_{t+s}g} - e^{KtH_tg}}{s}\right) d\nu \;=\; -\,Kt \int_M \frac{|\nabla^- H_tg|^2}{2}\, e^{KtH_tg}\, d\nu. \tag{22.37}$$

In summary, for any $t > 0$, ϕ is right-differentiable at t and

$$\frac{d^+\phi(t)}{dt} := \lim_{s\to 0^+} \left[\frac{\phi(t+s) - \phi(t)}{s}\right] \tag{22.38}$$

$$= \frac{1}{Kt^2 \displaystyle\int_M e^{KtH_tg}\, d\nu} \left[-\left(\int_M e^{KtH_tg}\, d\nu\right) \log\left(\int_M e^{KtH_tg}\, d\nu\right) \right.$$

$$\left. +\; \int_M (KtH_tg)\, e^{KtH_tg}\, d\nu \;-\; \frac{1}{2K} \int_M \big(Kt|\nabla^- H_tg|\big)^2\, e^{KtH_tg}\, d\nu \right]. \tag{22.39}$$

Because ν satisfies a logarithmic Sobolev inequality with constant K, the quantity inside square brackets is nonpositive. So ϕ is nonincreasing and the proof is complete. □

Before going to the proof of Theorem 22.15(ii), it might be a good idea to think over the next exercise, so as to understand more "concretely" why Talagrand inequalities are related to Poincaré inequalities.

Exercise 22.18. Use Otto's calculus to show that, at least formally,

$$\|h\|_{H^{-1}(\nu)} = \lim_{\varepsilon\to 0} \frac{W_2\big((1+\varepsilon h)\nu,\, \nu\big)}{\varepsilon},$$

where h is smooth and bounded (and compactly supported, if you wish), $\int h\, d\nu = 0$, and the dual Sobolev norm $H^{-1}(\nu)$ is defined by

$$\|h\|_{H^{-1}(\nu)} \;=\; \sup_{h\neq 0} \frac{\|h\|_{L^2(\nu)}}{\|\nabla h\|_{L^2(\nu)}} \;=\; \big\|\nabla(L^{-1}h)\big\|_{L^2(\nu)},$$

where as before $L = \Delta - \nabla V \cdot \nabla$. Deduce that, at least formally, the Talagrand inequality reduces, in the limit when $\mu = (1+\varepsilon h)\,\nu$ and $\varepsilon \to 0$, to the **dual Poincaré inequality**

$$\left[\int h\, d\nu = 0\right] \;\Longrightarrow\; \qquad \|h\|_{H^{-1}(\nu)} \le \frac{\|h\|_{L^2(\nu)}}{\sqrt{K}}.$$

Proof of Theorem 22.15, part (ii). Let $h : M \to \mathbb{R}$ be a bounded Lipschitz function satisfying $\int_M h\, d\nu = 0$. Introduce

$$\psi(t) = \int_M e^{KtH_t h}\, d\nu. \tag{22.40}$$

From the dual formulation of Talagrand's inequality (Proposition 22.2 for $p = 2$), $\psi(t)$ is bounded above by $\exp(Kt \int_M h\, d\nu) = 1$; hence ψ has a maximum at $t = 0$. Combining this with $\int h\, d\nu = 0$, we find

$$0 \le \limsup_{t \to 0^+} \left(\frac{1 - \psi(t)}{Kt^2} \right) = \limsup_{t \to 0^+} \int_M \left(\frac{1 + Kt\, h - e^{K_t\, H_t h}}{Kt^2} \right) d\nu. \tag{22.41}$$

By the boundedness of $H_t h$ and Proposition 22.13(iv),

$$\begin{aligned} e^{KtH_t h} &= 1 + KtH_t h + \frac{K^2 t^2}{2} (H_t h)^2 + O(t^3) \\ &= 1 + KtH_t h + \frac{K^2 t^2}{2} h^2 + o(t^2). \end{aligned} \tag{22.42}$$

So the right-hand side of (22.41) equals

$$\limsup_{t \to 0^+} \int_M \left(\frac{h - H_t h}{t} \right) d\nu - \frac{K}{2} \int_M h^2\, d\nu. \tag{22.43}$$

By Proposition 22.13(v), $(h - H_t h)/t$ is bounded; so we can apply Fatou's lemma, in the form

$$\limsup_{t \to 0^+} \int_M \left(\frac{h - H_t h}{t} \right) d\nu \le \int_M \limsup_{t \to 0^+} \left(\frac{h - H_t h}{t} \right) d\nu. \tag{22.44}$$

Then Proposition 22.13(vi) implies that

$$\int_M \limsup_{t \to 0^+} \left(\frac{h - H_t h}{t} \right) d\nu \le \int_M \frac{|\nabla^- h|^2}{2}\, d\nu. \tag{22.45}$$

All in all, the right-hand side of (22.41) can be bounded above by

$$\frac{1}{2} \int_M |\nabla^- h|^2\, d\nu - \frac{K}{2} \int_M h^2\, d\nu. \tag{22.46}$$

So (22.46) is always nonnegative, which concludes the proof of the Poincaré inequality. □

To close this section, I will show that the Talagrand inequality does imply a logarithmic Sobolev inequality under certain curvature assumptions.

Theorem 22.19 (T_2 sometimes implies log Sobolev). *Let M be a Riemannian manifold and let $\nu = e^{-V} \mathrm{vol} \in P_2(M)$ a reference measure on M, $V \in C^2(M)$. Assume that ν satisfies a Talagrand inequality $T_2(\lambda)$, and a curvature-dimension inequality $\mathrm{CD}(K, \infty)$ for some $K > -\lambda$. Then ν also satisfies a logarithmic Sobolev inequality with constant*

$$\widetilde{\lambda} = \max \left[\frac{\lambda}{4} \left(1 + \frac{K}{\lambda} \right)^2,\ K \right].$$

Proof of Theorem 22.19. From the assumptions and Theorem 20.7, the nonnegative quantities $H = H_\nu(\mu)$, $W = W_2(\mu, \nu)$ and $I = I_\nu(\mu)$ satisfy the inequalities

$$H \le W \sqrt{I} - \frac{\lambda W^2}{2}, \qquad W \le \sqrt{\frac{2H}{K}}.$$

It follows by an elementary calculation that $H \le I/(2\widetilde{\lambda})$, so ν satisfies a logarithmic Sobolev inequality with constant $\widetilde{\lambda}$. □

Poincaré inequalities and quadratic-linear transport cost

So far we have encountered transport inequalities involving the quadratic cost function $c(x,y) = d(x,y)^2$, and the linear cost function $c(x,y) = d(x,y)$. Remarkably, Poincaré inequalities can be recast in terms of transport cost inequalities involving a cost function which behaves quadratically for small distances, and linearly for large distances. As discovered by Bobkov and Ledoux, they can also be rewritten as **modified logarithmic Sobolev inequalities**, which are just usual logarithmic Sobolev inequalities, except that there is a Lipschitz constraint on the logarithm of the density of the measure. These two reformulations of Poincaré inequalities will be discussed below.

Definition 22.20 (quadratic-linear cost). *Let $(\mathcal{X}, d)$ be a metric space. The quadratic-linear cost $c_{q\ell}$ on $\mathcal{X}$ is defined by*

$$c_{q\ell}(x,y) = \begin{cases} d(x,y)^2 & \text{if } d(x,y) \le 1; \\ d(x,y) & \text{if } d(x,y) > 1. \end{cases}$$

In a compact writing, $c_{q\ell}(x,y) = \max(d(x,y)^2, d(x,y))$. The optimal total cost associated with $c_{q\ell}$ will be denoted by $C_{q\ell}$.

Theorem 22.21 (Reformulations of Poincaré inequalities). *Let M be a Riemannian manifold equipped with a reference probability measure ν. Then the following three statements are equivalent:*

(i) ν satisfies a Poincaré inequality;

(ii) There are constants $c > 0$, $K > 0$ such that for any Lipschitz probability density ρ,

$$|\nabla \log \rho| \le c \implies U_\nu(\mu) \le \frac{I_\nu(\mu)}{K}, \qquad \mu = \rho\,\nu. \tag{22.47}$$

(iii) $\nu \in P_1(M)$ and there is a constant $C > 0$ such that

$$\forall \mu \in P_1(M), \qquad C_{q\ell}(\mu,\nu) \le C\, H_\nu(\mu). \tag{22.48}$$

Remark 22.22. The equivalence between (i) and (ii) can be made more precise. As the proof shows, if ν satisfies a Poincaré inequality with constant λ, then for any $c < 2\sqrt{\lambda}$ there is an explicit constant $K = K(c) > 0$ such that (22.47) holds true; and the constant $K(c)$ converges to λ as $c \to 0$. Conversely, if for each $c > 0$ we call $K(c)$ the best constant in (22.47), then ν satisfies a Poincaré inequality with constant $\lambda = \lim_{c\to 0} K(c)$.

Remark 22.23. The equivalence between (i) and (ii) remains true when the Riemannian manifold M is replaced by a general metric space. On the other hand, the equivalence with (iii) uses at least a little bit of the Riemannian structure.

Theorem 22.21 will be obtained as a consequence of several propositions which have their own interest. The first one is about the behavior of the Hamilton–Jacobi semigroup, when the quadratic Lagrangian $L(c) = c^2/2$ has been replaced by a more general function of the speed.

Proposition 22.24 (Properties of the general Hamilton–Jacobi semigroup). *Let $L : \mathbb{R}_+ \to \mathbb{R}_+$ be a strictly increasing convex continuous function with $L(0) = 0$, and let L^* be its Legendre transform. If $L'(\infty) < +\infty$, further assume that $L^*(L'(\infty)) < +\infty$. Let M*

be a Riemannian manifold equipped with its geodesic distance. For any bounded Lipschitz continuous function $f : M \to \mathbb{R}$ define the evolution $(H_t\varphi)_{t\geq 0}$ by

$$\begin{cases} H_0\,\varphi = \varphi, \\ \\ (H_t\,\varphi)(x) = \inf\limits_{y\in M} \left[\varphi(x) + t\,L\left(\dfrac{d(x,y)}{t}\right)\right] \qquad (t>0,\ x\in M). \end{cases} \tag{22.49}$$

Then

(i) For all $s,t \geq 0$, $H_t H_s f = H_{t+s} f$.

(ii) For all $x \in M$, $\inf f \leq (H_t f)(x) \leq f(x)$; moreover, the infimum in (22.49) can be restricted to $y \in B(x, R(f,t))$, where

$$R(f,t) = t\,L^{-1}\left(\frac{\sup f - \inf f}{t}\right).$$

(iii) For all $t>0$, $H_t f$ is locally Lipschitz on M, and $\|H_t f\|_{\mathrm{Lip}} \leq L'(\infty)$; in particular, $R(H_t f, s) < +\infty$ for all $s>0$.

(iv) For all $x\in M$, $(H_t f)(x)$ is a nonincreasing function of t. Moreover, for any $t>0$, $(H_{t+s}f)$ converges monotonically, and locally uniformly, to $H_t f$ as $s\to 0$.

(v) For all $t\geq 0$, $s>0$, $x\in M$,

$$\frac{|H_{t+s}f(x) - H_t f(x)|}{s} \leq L^*\Big(\|H_t f\|_{\mathrm{Lip}(B(x,R(s)))}\Big).$$

(vi) For all $x\in M$ and $t\geq 0$,

$$\liminf_{s\to 0^+} \frac{(H_{t+s}f)(x) - (H_t f)(x)}{s} \geq -\,L^*(|\nabla^- H_t f|). \tag{22.50}$$

(vii) For all $x\in M$ and $t>0$,

$$\lim_{s\to 0^+} \frac{(H_{t+s}f)(x) - (H_t f)(x)}{s} = -\,L^*(|\nabla^- H_t f|). \tag{22.51}$$

The proof of this proposition is omitted; it is quite similar to the proof of Proposition 22.13.

The second result is a generalisation of Theorem 22.15.

Theorem 22.25 (From generalized logarithmic Sobolev to transport to generalized Poincaré). *Let M be a Riemannian manifold equipped with its geodesic distance d and a reference measure $\nu \in P_2(M)$. Let L satisfy the same assumptions as in Proposition 22.24, let $c_L(x,y) = L(d(x,y))$ and let C_L be the optimal transport cost associated with the cost function c. Assume further that $L(r) \leq C(1+r)^p$ for some $p\in[1,2]$ and some $C>0$. Then*

(i) Further assume that $L^*(ts) \leq t^2 L^*(s)$ for all $t\in[0,1]$, $s\geq 0$. If there is $\lambda\in(0,1]$ such that ν satisfies the generalized logarithmic Sobolev inequality with constant λ:

$$\forall \mu \in P(M), \qquad H_\nu(\mu) \leq \frac{1}{\lambda}\int L^*\big(|\nabla^-\log\rho|\big)\,d\mu, \qquad \mu = \rho\,\nu, \quad \log\rho \in \mathrm{Lip}(M);$$

then ν also satisfies the following transport inequality:

$$\forall \mu \in P_p(M), \qquad C_L(\mu,\nu) \le \frac{H_\nu(\mu)}{\lambda}. \tag{22.52}$$

(ii) If ν satisfies (22.52)*, then it also satisfies the generalized Poincaré inequality with constant λ:*

$$\forall f \in \mathrm{Lip}(M),\ \|f\|_{\mathrm{Lip}} \le L'(\infty), \quad \int f\,d\nu = 0 \Longrightarrow \quad \int f^2\,d\nu \le \frac{2}{\lambda}\int L^*(|\nabla^- f|)\,d\nu.$$

Proof of Theorem 22.25. The proof is exactly similar to the proof of Theorem 22.15. After picking up $g \in C_b(M)$, one introduces the function

$$\phi(t) = \frac{1}{\lambda t}\log \int e^{\lambda t\, H_t g}\,d\nu.$$

Now it is not always true that ϕ is continuous at $t = 0$, but at least the monotonicity of $H_t g$ implies that

$$\lim_{t\to 0^+} \phi(t) \le \phi(0).$$

Then one can compute the right derivative

$$\begin{aligned}
\frac{d^+\phi(t)}{dt} &:= \lim_{s\to 0^+}\left[\frac{\phi(t+s)-\phi(t)}{s}\right] \qquad (22.53)\\
&= \frac{1}{\lambda t^2\displaystyle\int_M e^{\lambda t H_t g}\,d\nu}\left[-\left(\int_M e^{\lambda t H_t g}\,d\nu\right)\log\left(\int_M e^{\lambda t H_t g}\,d\nu\right)\right.\\
&\qquad \left. +\int_M (\lambda t H_t g) e^{\lambda t H_t g}\,d\nu - \frac{1}{2\lambda}\int_M \left(\lambda^2 t^2 L^*(|\nabla^- H_t g|)\right) e^{\lambda t H_t g}\,d\nu\right]\\
&= \frac{1}{\lambda t^2\displaystyle\int_M e^{\lambda t H_t g}\,d\nu}\left[-\left(\int_M e^{\lambda t H_t g}\,d\nu\right)\log\left(\int_M e^{\lambda t H_t g}\,d\nu\right)\right.\\
&\qquad \left. +\int_M (\lambda t H_t g) e^{\lambda t H_t g}\,d\nu - \frac{1}{2\lambda}\int_M L^*\left(\lambda t|\nabla^- H_t g|\right) e^{\lambda t H_t g}\,d\nu\right], \qquad (22.54)
\end{aligned}$$

where the inequality $L^*(\lambda t s) \le \lambda^2 t^2 L^*(s)$ was used. By assumption, the quantity inside square brackets is nonpositive, so ϕ is nonincreasing on $(0,1]$, and therefore on $[0,1]$. The inequality $\phi(1) \le \phi(0)$ can be recast as

$$\frac{1}{\lambda}\log \int_M e^{\lambda\, \inf_{y\in M}\left[g(y)+L(d(x,y))\right]}\,\nu(dx) \le \int_M g\,d\nu,$$

which by Theorem 5.21 is the dual formulation of (22.52).

Part (ii) of the theorem is similar to part (ii) of Theorem 22.15. □

Now we have enough tools at our disposal to carry on the proof of Theorem 22.21.

Proof of Theorem 22.21. We start by the proof of (i) ⇒ (ii). Let $f = \log\rho - \int(\log\rho)\,d\nu$; so $\int f\,d\nu = 0$ and the assumption in (ii) reads $|\nabla f| \le c$. Moreover, with $a = \int(\log\rho)\,d\nu$ and $X = \int e^f\,d\nu$,

$$I_\nu(\mu) = e^a \int |\nabla f|^2\, e^f\,d\nu;$$

$$\begin{aligned} H_\nu(\mu) &= \int (f+a)e^{f+a}\,d\nu - \left(\int e^{f+a}\,d\nu\right)\log\left(\int e^{f+a}\,d\nu\right) \\ &= e^a\left(\int fe^f\,d\nu - \int e^f\,d\nu + 1\right) - e^a(X\log X - X + 1) \\ &\le e^a\left(\int fe^f\,d\nu - \int e^f\,d\nu + 1\right). \end{aligned}$$

So it is sufficient to prove

$$|\nabla f| \le c \Longrightarrow \qquad \int \left(fe^f - e^f + 1\right)d\nu \le \frac{1}{K}\int |\nabla f|^2 e^f\,d\nu. \tag{22.55}$$

In the sequel, c is any constant satisfying $0 < c < 2\sqrt{\lambda}$, and we shall keep track of the dependence of K on c. Inequality (22.55) will be proven in two steps:

$$\int f^2\,d\nu \le e^{c\sqrt{5}/\lambda}\int f^2 e^{-|f|}\,d\nu. \tag{22.56}$$

$$\int f^2 e^f\,d\nu \le \frac{1}{\lambda}\left(\frac{2\sqrt{\lambda}+c}{2\sqrt{\lambda}-c}\right)^2\int |\nabla f|^2 e^f\,d\nu; \tag{22.57}$$

Note that the bound on $|\nabla f|$ is crucial in both inequalities.

Once (22.56) and (22.57) are established, the result follows immediately. Indeed, the right-hand side of (22.56) is obviously bounded by the left-hand side of (22.57), so both expressions are bounded above by a constant multiple of $\int |\nabla f|^2 e^f\,d\nu$. On the other hand, an elementary study shows that

$$\forall f \in \mathbb{R}, \qquad fe^f - e^f + 1 \le \max\left(f^2, f^2 e^f\right),$$

so (22.55) holds true.

To obtain (22.56), we proceed as follows. The elementary inequality $2|f|^3 \le \delta f^2 + \delta^{-1} f^4$ ($\delta > 0$) integrates up to

$$\begin{aligned} 2\int |f|^3\,d\nu &\le \delta\int f^2\,d\nu + \delta^{-1}\int f^4\,d\nu \\ &= \delta\int f^2\,d\nu + \delta^{-1}\left(\int f^2\,d\nu\right)^2 + \left[\int (f^2)^2\,d\nu - \left(\int f^2\,d\nu\right)^2\right]. \end{aligned} \tag{22.58}$$

By Poincaré inequality, $\int f^2\,d\nu \le (1/\lambda)\int |\nabla f|^2\,d\nu \le c^2/\lambda$, so $(\int f^2\,d\nu)^2 \le (c^2/\lambda)\int f^2\,d\nu$. Also by Poincaré inequality,

$$\int (f^2)^2\,d\nu - \left(\int f^2\,d\nu\right)^2 \le (1/\lambda)\int |\nabla(f^2)|^2\,d\nu = (4/\lambda)\int f^2|\nabla f|^2\,d\nu \le (4c^2/\lambda)\int f^2\,d\nu.$$

Plugging this information back in (22.58), we obtain

$$2\int |f|^3\,d\nu \le \left(\delta + \frac{5c^2}{\delta\lambda}\right)\int f^2\,d\nu.$$

The choice $\delta = \sqrt{5c^2/\lambda}$ yields

$$\int |f|^3\,d\nu \le c\sqrt{\frac{5}{\lambda}}\int f^2\,d\nu. \tag{22.59}$$

Then by Jensen's inequality, applied with the convex function $x \to e^{-|x|}$ and the probability measure $\sigma = f^2\,\nu/(\int f^2\,d\nu)$, we get

$$\int f^2 e^{-|f|}\,d\nu = \left(\int e^{-|f|}\,d\sigma\right)\left(\int f^2\,d\nu\right) \geq e^{-\int |f|\,d\sigma}\left(\int f^2\,d\nu\right),$$

or in other words

$$\int f^2\,d\nu \leq \exp\left(\frac{\int |f|^3\,d\nu}{\int f^2\,d\nu}\right)\int f^2 e^{-|f|}\,d\nu.$$

Combining this inequality with (22.59) finishes the proof of (22.56).

To establish (22.57), we first use the condition $\int f\,d\nu = 0$ and the Poincaré inequality to write

$$\begin{aligned}
\left(\int f e^{f/2}\,d\nu\right)^2 &= \frac{1}{4}\left(\int [f(x)-f(y)]\,[e^{f(x)/2}-e^{f(y)/2}]\,d\nu(x)\,d\nu(y)\right)^2 \\
&\leq \frac{1}{4}\left(\int |f(x)-f(y)|^2\,d\nu(x)\,d\nu(y)\right)\left(\int [e^{f(x)/2}-e^{f(y)/2}]^2\,d\nu(x)\,d\nu(y)\right) \\
&= \left(\int f^2\,d\nu - \left(\int f\,d\nu\right)^2\right)\left(\int e^f\,d\nu - \left(\int e^{f/2}\,d\nu\right)^2\right) \\
&\leq \frac{1}{\lambda^2}\left(\int |\nabla f|^2\,d\nu\right)\left(\int |\nabla(e^{f/2})|^2\,d\nu\right) \\
&= \frac{c^2}{4\lambda^2}\int |\nabla f|^2 e^f\,d\nu. \qquad (22.60)
\end{aligned}$$

Next, also by the Poincaré inequality and the chain-rule,

$$\begin{aligned}
\int f^2 e^f\,d\nu - \left(\int f e^{f/2}\,d\nu\right)^2 &\leq \frac{1}{\lambda}\int \left|\nabla(f e^{f/2})\right|^2 d\nu \qquad (22.61) \\
&= \frac{1}{\lambda}\int |\nabla f|^2\left(1+\frac{f}{2}\right)^2 e^f\,d\nu \\
&= \frac{1}{\lambda}\left(\int |\nabla f|^2 e^f\,d\nu + \int |\nabla f|^2 f e^f\,d\nu + \frac{1}{4}\int |\nabla f|^2 f^2\,e^f\,d\nu\right) \\
&\leq \frac{1}{\lambda}\left(\int |\nabla f|^2 e^f\,d\nu + c\sqrt{\int |\nabla f|^2 e^f\,d\nu}\,\sqrt{\int f^2 e^f\,d\nu} + \frac{c^2}{4}\int f^2 e^f\,d\nu\right). \qquad (22.62)
\end{aligned}$$

By adding up (22.60) and (22.62), we obtain

$$\int f^2 e^f\,d\nu \leq \left(\frac{1}{\lambda}+\frac{c^2}{4\lambda^2}\right)\int |\nabla f|^2 e^f\,d\nu + \frac{c}{\lambda}\sqrt{\int |\nabla f|^2 e^f\,d\nu}\,\sqrt{\int f^2 e^f\,d\nu} + \frac{c^2}{4\lambda}\int f^2 e^f\,d\nu.$$

This inequality involving the two quantities $\int f^2 e^f\,d\nu$ and $\int |\nabla f|^2 e^f\,d\nu$ can be transformed into (22.57). (Here the fact that $c^2/(4\lambda) < 1$ is crucial.) This complete the proof of (i) $\Rightarrow$ (ii).

Now we shall see that (ii) $\Rightarrow$ (iii). Let ν satisfy a modified logarithmic Sobolev inequality as in (22.47). Let then $L(s) = cs^2/2$ for $0 \leq s \leq 1$, $L(s) = c(s-1/2)$ for $s > 1$. The function L so defined is convex, strictly increasing, with $L'(\infty) = c$. Its Legendre transform L^* is quadratic on $[0, c]$ and identically $+\infty$ on $(c, +\infty)$. So (**??**) can be rewritten

$$H_\nu(\mu) \le \frac{2c}{K} \int L^*(|\nabla \log \rho|)\, d\mu$$

Since $L^*(tr) \le t^2\, L^*(r)$ for all $t \in [0,1]$, $r \ge 0$, we can apply Theorem 22.25(i) to deduce the modified transport inequality

$$C_L(\mu,\nu) \le \frac{2c}{K} H_\nu(\mu), \tag{22.63}$$

which is easily seen to be equivalent to (iii).

It remains to check (iii) $\Rightarrow$ (i). If ν satisfies (iii), or equivalently (22.63), then it also satisfies the generalized Poincaré inequality of Theorem 22.25(ii). Pick up any Lipschitz function f and apply this inequality to εf, where ε is small enough that $\varepsilon \|f\|_{\mathrm{Lip}} < c$; the result is

$$\int f\, d\nu = 0 \Longrightarrow \quad \varepsilon^2 \int f^2\, d\nu \le \frac{2}{\lambda} \int L^*(\varepsilon |\nabla^- f|)\, d\nu.$$

Since L^* is quadratic on $[0,c]$, factors ε^2 cancel out on both sides, and we are back with the usual Poincaré inequality. □

Example 22.26. Prove directly the implication (ii) $\Rightarrow$ (i).

Let us now see the implications of Theorem 22.21 in terms of concentration of measure.

Theorem 22.27 (Concentration of measure from Poincaré inequality). *Let M be a Riemannian manifold equipped with its geodesic distance, and with a reference probability measure ν. Assume that ν satisfies a Poincaré inequality with constant λ. Then there is a constant $C = C(\lambda) > 0$ such that for any Borel set A,*

$$\forall r \ge 0, \qquad \nu[A^r] \ge 1 - \frac{e^{-C\,\min(r,r^2)}}{\nu[A]}. \tag{22.64}$$

Moreover, for any $f \in \mathrm{Lip}(M)$ (resp. $\mathrm{Lip}(M) \cap L^1(\nu)$),

$$\nu\Big[\big\{x;\ f(x) \ge m + r\big\}\Big] \le e^{-C\,\min\left(\frac{r}{\|f\|_{\mathrm{Lip}}}, \frac{r^2}{\|f\|^2_{\mathrm{Lip}}}\right)}, \tag{22.65}$$

where m is a median of f (resp. the mean value of f).

Proof of Theorem 22.27. The proof of (22.64) is similar to the implication (i) $\Rightarrow$ (iii) in Theorem 22.9. Define $B = M \setminus A^r$, and let $\nu_A = (1_A)\nu/\nu[A]$, $\nu_B = (1_B)\nu/\nu[B]$. Obviously, $C_{q\ell}(\nu_A,\nu_B) \ge \min(r,r^2)$. The elementary inequality $\min(a+b,(a+b)^2) \le 4[\min(a,a^2) + \min(b,b^2)]$ implies $C_{q\ell}(\nu_A,\nu_B) \le 4[C_{q\ell}(\nu_A,\nu) + C_{q\ell}(\nu_B,\nu)]$. So

$$\min(r,r^2) \le 4[C_{q\ell}(\nu_A,\nu) + C_{q\ell}(\nu_B,\nu)].$$

Since ν satisfies (22.48), there is a constant C such that

$$\begin{aligned}\min(r,r^2) &\le C\big(H_\nu(\nu_A) + H_\nu(\nu_B)\big)\\ &= C\Big(\log\frac{1}{\nu[A]} + \log\frac{1}{1-\nu[A^r]}\Big).\end{aligned}$$

Then (22.64) follows immediately. Then (**??**) is obtained by arguments similar to those used before in the proof of Theorem 22.9. □

Example 22.28. The exponential measure $\nu(dx) = (1/2)e^{-|x|}\,dx$ does not admit Gaussian tails, so it fails to satisfy properties of Gaussian concentration expressed in Theorem **??**. However, it does satisfy a Poincaré inequality. So (22.64), (22.65) hold true for this measure.

Consider now the problem of concentration of measure in a *product space*, say $(M^N, \nu^{\otimes N})$, where ν satisfies a Poincaré inequality. We may equip M^N with the metric

$$d_2(x,y) = \sqrt{\sum_i d(x_i,y_i)^2};$$

then $\mu^{\otimes N}$ will satisfy a Poincaré inequality with the same constant as ν, and we may apply Theorem 22.27 to study concentration in $(M^N, d_2, \nu^{\otimes N})$. There is however a more interesting approach, due to Talagrand, in which one uses both the distance d_2 and the distance

$$d_1(x,y) = \sum_i d(x_i,y_i).$$

The procedure is as follows: Given a Borel set $A \subset M^N$, first enlarge it by r in distance d_2 (that is, consider all points which lie at a distance less than r from A); then enlarge the result by r^2 in distance d_1. This is explained in the next theorem, where $A^{r;d}$ stands for the enlargement of A by r in distance d, and $\|f\|_{\mathrm{Lip}}(\mathcal{X}, d)$ stands for the Lipschitz norm of f on $\mathcal{X}$ with respect to the distance d.

Theorem 22.29 (Concentration in product spaces from Poincaré inequalities). *Let M be a Riemannian manifold equipped with its geodesic distance d, and with a reference probability measure ν. Assume that ν satisfies a Poincaré inequality with constant λ. Then there is a constant $C = C(\lambda)$ such that for all $N \in \mathbb{N}$, and for any Borel set $A \subset M^N$,*

$$\nu^{\otimes N}[A] \geq \frac{1}{2} \quad \Longrightarrow \quad \nu^{\otimes N}\Big[(A^{r;d_2})^{r^2;d_1}\Big] \geq 1 - e^{-C\,r^2}. \tag{22.66}$$

Moreover, for any $f \in \mathrm{Lip}(M^N, d_1) \cap \mathrm{Lip}(M^N, d_2)$ (resp. $\mathrm{Lip}(M^N, d_1) \cap \mathrm{Lip}(M^N, d_2) \cap L^1(\nu^{\otimes} N)$),

$$\nu^{\otimes N}\Big[\big\{x;\ f(x) \geq m + r\big\}\Big] \leq e^{-C\,\min\left(\frac{r}{\|f\|_{\mathrm{Lip}(M^N,d_1)}}, \frac{r^2}{\|f\|^2_{\mathrm{Lip}(M^N,d_2)}}\right)}, \tag{22.67}$$

where m is a median of f (resp. the mean value of f) with respect to the measure $\nu^{\otimes N}$.

Proof of Theorem 22.29. Once again, the implication (22.66) $\Rightarrow$ (22.67) follows arguments similar to those used in the proof of Theorem 22.9 (actually these two statements are equivalent, up to a loss of constants); so we concentrate on the proof of (22.66).

By Theorem 22.21, ν satisfies a transport-cost inequality of the form

$$\forall \mu \in P_1(M), \quad C_{q\ell}(\mu,\nu) \leq C\,H_\nu(\mu).$$

On M^N define the cost

$$c(x,y) = \sum c_{q\ell}(x_i,y_i),$$

and let C be the associated optimal cost functional.

By Remark 22.7, $\nu^{\otimes N}$ satisfies an inequality of the form

$$\forall \mu \in P_1(M^N), \quad C(\mu,\nu^{\otimes N}) \leq C\,H_{\nu^{\otimes N}}(\mu). \tag{22.68}$$

Let A be a Borel set of M^N with $\nu^{\otimes N}[A] \geq 1/2$, and let $r > 0$ be given. Let $B = M^N \setminus (A^{r;d_2})^{r^2;d_1}$. Let ν_B be obtained by conditioning ν on B (that is, $\nu_B = (1_B)\nu/\nu[B]$). Consider the problem of transporting ν_B to ν optimally, with the cost c. At least a portion $\nu^{\otimes N}[A] \geq 1/2$ of the mass has to go to from B to A, so

$$C(\nu_B, \nu^{\otimes N}) \geq \frac{1}{2} \inf_{x \in A,\, y \in B} c(x,y) =: \frac{1}{2}\, c(A,B).$$

On the other hand, by (22.68),

$$C(\mu, \nu^{\otimes N}) \leq C\, H_{\nu^{\otimes N}}(\mu) = C\, \log \frac{1}{\nu[B]}.$$

By combining these two inequalities, we get

$$\nu\Big[(A^{r;d_2})^{r^2;d_1}\Big] = 1 - \nu[B] \geq 1 - e^{-1/(2C)\, c(A,B)}.$$

To prove (22.66), it only remains to check that

$$c(A,B) \geq r^2.$$

So let $x = (x_1, \ldots, x_N) \in A$, and let $y \in M^N$ such that $c(x,y) < r^2$; the goal is to show that $y \in (A^{r;d_2})^{r^2;d_1}$. For each $i \in \{1, \ldots, N\}$, define $z_i = x_i$ if $d(x_i, y_i) > 1$, $z_i = y_i$ otherwise. Then

$$d_2(x,z)^2 = \sum_{d(x_i,y_i) \leq 1} d(x_i,y_i)^2 \leq \sum_i c_{q\ell}(x_i,y_i) = c(x,y) < r^2;$$

so $z \in A^{r;d_2}$. Similarly,

$$d_1(z,y) = \sum_{d(x_i,y_i) > 1} d(x_i,y_i) \leq \sum_i c_{q\ell}(x_i,y_i) = c(x,y) < r^2;$$

so y lies at a distance at most r^2 from z, in distance d_1. This concludes the proof. □

Example 22.30. Let $\nu(dx)$ be the exponential measure $e^{-|x|}dx/2$ on $\mathbb{R}$, then $\nu^{\otimes N}(dx) = (1/2^N)e^{-\sum |x_i|} \prod dx_i$ on $\mathbb{R}^N$. Theorem 22.29 shows that for every Borel set $A \subset \mathbb{R}^N$ with $\nu^{\otimes N}[A] \geq 1/2$ and any $\delta > 0$,

$$\nu^{\otimes N}\big[A + B_r^{d_2} + B_{r^2}^{d_1}\big] \geq 1 - e^{-cr^2} \tag{22.69}$$

where B_r^d stands for the ball of center 0 and radius r in $\mathbb{R}^N$ for the distance d.

Remark 22.31. Strange as this may seem, inequality (22.69) contains (up to numerical constants) the Gaussian concentration of the Gaussian measure! Let indeed $T : \mathbb{R} \to \mathbb{R}$ be the increasing rearrangement of the exponential measure ν onto the one-dimensional Gaussian measure γ (so $T_\# \nu = \gamma$, $(T^{-1})_\# \gamma = \nu$). An explicit computation shows that

$$|T(x) - T(y)| \leq C \min(|x-y|, \sqrt{|x-y|}) \tag{22.70}$$

for some numeric constant C. Let then $T_N(x_1, \ldots, x_N) = (T(x_1), \ldots, T(x_N))$; obviously $(T_N)_\#(\nu^{\otimes N}) = \gamma^{\otimes N}$, $(T_N)_\#^{-1}(\gamma^{\otimes N}) = \nu^{\otimes N}$. Let A be any Borel set, and let $y \in T_N^{-1}(A) + B_r^{d_2} + B_{r^2}^{d_1}$. This means that there are w and x such that $T_N(w) \in A$, $|x - w|_2 \leq r$, $|y - x|_1 \leq r^2$. Then by (22.70),

$$
\begin{aligned}
|T_N(w) - T_N(y)|_2^2 &= \sum |T(w_i) - T(y_i)|^2 \\
&\le C^2 \sum_i \min\big(|w_i - y_i|, |w_i - y_i|^2\big) \\
&\le C^2 \Big(\sum_{|w_i - x_i| \ge |x_i - y_i|} 2|w_i - x_i| \; + \sum_{|w_i - x_i| < |x_i - y_i|} 4|x_i - y_i|^2 \Big) \\
&\le 4C^2 \big(\sum |x_i - w_i| + \sum |x_i - y_i|^2 \big) \; \le 8C^2 r^2.
\end{aligned}
$$

This means that $T_N(y) \in A + B^{d_2}_{\sqrt{8}Cr}$. In summary, if $C' = \sqrt{8}C$, then

$$
T_N\big(T_N^{-1}(A) + B^{d_2}_r + B^{d_1}_{r^2}\big) \subset A + B^{d_2}_{C'r}.
$$

As a consequence, if $A \subset \mathbb{R}^N$ is such that $\gamma^{\otimes N}[A] \ge 1/2$, then $\nu^{\otimes N}[T_N^{-1}(A)] = \gamma^{\otimes N}[A] \ge 1/2$, and

$$
\begin{aligned}
\gamma^{\otimes N}[A^{C'r}] &\ge \gamma^{\otimes N}\Big[T_N\big(T_N^{-1}(A) + B^{d_2}_r + B^{d_1}_{r^2}\big)\Big] \\
&= \nu^{\otimes N}\big[T_N^{-1}(A) + B^{d_2}_r + B^{d_1}_{r^2}\big] \\
&\ge 1 - e^{-cr^2}
\end{aligned}
$$

for some numeric constant $c > 0$. This is precisely the Gaussian concentration property appearing in Theorem 22.9(iii).

Remark 22.32. In certain situations, (22.69) provides sharper concentration properties for the Gaussian measure, than the usual Gaussian concentration bounds. This might look paradoxical, but can be explained by the fact that Gaussian concentration considers *arbitrary* sets A, while in many problems one is led to study the concentration of measure around certain very particular sets, for instance with a "cubic" structure; then inequality (22.69) might be very efficient.

Example 22.33. Let $A = \{x \in \mathbb{R}^N;\ \max |x_i| \le m\}$ be the centered cube of side $2m$, where $m = m(N) \to \infty$ is chosen in such a way that $\gamma^{\otimes N}[A] \ge 1/2$. (It is a classical fact that $m = O(\sqrt{\log N})$, but we don't need that information.) If $r \ge 1$ is small with respect to m, then the enlargement of the cube is dominated by the behavior of T close to $T^{-1}(m)$. Since $T(x)$ behaves approximately like $\sqrt{x}$ for large values of x, $T^{-1}(m)$ is of the order m^2; and close to m^2 the Lipschitz norm of T is $O(1/m)$. Then the computation before can be sharpened into

$$
T_N\big(T_N^{-1}(A) + B^{d_2}_r + B^{d_1}_{r^2}\big) \subset A + B^{d_2}_{C'r^2/m}.
$$

So the concentration of measure can be felt by enlarging A by a distance of the order of $r^2/m \ll r$.

Dimension-dependent inequalities

There is no well-identified analogue of Talagrand inequalities that would take advantage of the finiteness of the dimension to provide sharper concentration inequalities. In this section I shall suggest some natural possibilities, focusing on positive curvature for simplicity.

Theorem 22.34 (Dimension-dependent transport-energy inequalities). *Let M be a Riemannian manifold equipped with a reference measure $\nu = e^{-V}\mathrm{vol}$, $V \in C^2(M)$, satisfying the curvature-dimension bound $\mathrm{CD}(K,N)$ for some $K > 0$, $N \in (1,\infty)$. Then, for all $\mu = \rho\,\nu \in P_2(M)$,*

$$\int_M \left[N \left(\frac{\alpha}{\sin\alpha} \right)^{1-\frac{1}{N}} \rho(x_0)^{-\frac{1}{N}} - (N-1)\frac{\alpha}{\tan\alpha} \right] \pi(dx_0\,dx_1) \le 1, \tag{22.71}$$

where $\alpha(x_0,x_1) = \sqrt{K/(N-1)}\,d(x_0,x_1)$, and π is the unique optimal coupling between μ and ν. Equivalently,

$$\int_M \left[N \left(\frac{\alpha}{\sin\alpha} \right)^{1-\frac{1}{N}} - (N-1)\frac{\alpha}{\tan\alpha} - 1 \right] \pi(dx_0\,dx_1) \le \int \left(\frac{\alpha}{\sin\alpha} \right)^{1-\frac{1}{N}} \Big[(N-1)\rho - N\rho^{1-\frac{1}{N}} + 1 \Big]\, d\nu. \tag{22.72}$$

Remark 22.35. The function $(N-1)r - Nr^{1-\frac{1}{N}} + 1$ is nonnegative, and so is the integrand in the right-hand side of (22.72). If the coefficient $\alpha/\sin\alpha$ above would be replaced by 1, then the right-hand side of (22.72) would be just $\int [(N-1)\rho - N\rho^{1-\frac{1}{N}} + 1]\, d\nu = H_{N,\nu}(\rho)$.

Corollary 22.36 (Other forms of dimension-dependent transport-energy inequalities). *With the same assumptions and notation as in Theorem 22.34, the following inequalities hold true:*

$\forall p \in (1,\infty)$

$$\int \left[(Np-1) - (N-1)\frac{\alpha}{\tan\alpha} - N(p-1)\left(\frac{\sin\alpha}{\alpha} \right)^{\frac{1}{p-1}\left(1-\frac{1}{N}\right)} \right] d\pi \le H_{Np,\nu}(\mu); \tag{22.73}$$

$$\int \left[(2N-1) - (N-1)\frac{\alpha}{\tan\alpha} - N\exp\left(1 - \left(\frac{\alpha}{\sin\alpha} \right)^{1-\frac{1}{N}} \right) \right] d\pi \le 2H_{N,\nu}(\mu) - \int \rho^{1-\frac{1}{N}} \log\rho\, d\nu; \tag{22.74}$$

$$(N-1)\int \left(1 - \frac{\alpha}{\tan\alpha} + \log\frac{\alpha}{\sin\alpha} \right) d\pi \le H_{\infty,\nu}(\mu) \tag{22.75}$$

Proof of Theorem 22.34. Inequality (22.71) follows directly from Theorem 20.11 with $U(r) = -N(r^{1-1/N} - r)$ (recall Remark 20.12). To derive (22.72) from (22.71), it is sufficient to check that

$$\int NQ\, d\pi = \int Q\big[(N-1)\rho - 1\big]\, d\nu,$$

where $Q = (\alpha/\sin\alpha)^{1-\frac{1}{N}}$. But this is immediate because Q is a symmetric function of x_0 and x_1, and π has marginals $\mu = \rho\,\nu$ and ν, so

$$\int Q(x_0,x_1)\, d\nu(x_0) = \int Q(x_0,x_1)\, d\nu(x_1) = \int Q(x_0,x_1)\, d\pi(x_0,x_1) = \int Q(x_0,x_1)\,\rho(x_0)\, d\nu(x_0).$$

□

Proof of Corollary 22.36. Write again $Q = (\alpha/\sin\alpha)^{1-\frac{1}{N}}$. Then (22.73) follow from (22.71) upon use of the Young inequality $ab \le a^p/p + b^{p'}/p'$, where $p' = p/(p-1)$ is the conjugate exponent to p; indeed,

$$N p \rho^{1-\frac{1}{Np}} = (Np\rho) \left[\rho^{-\frac{1}{N}} Q\right]^{\frac{1}{p}} Q^{-\frac{1}{p}\left(1-\frac{1}{N}\right)} \le (Np\rho) \left[\frac{\rho^{-\frac{1}{N}} Q}{p} + \frac{Q^{-\frac{p'}{p}}}{p'}\right].$$

Then (22.74) and (22.75) are obtained by taking the limits $p \to 1$ and $p \to \infty$, respectively; or equivalently by applying the inequalities $ab \le a \log a - 2a + e^{b+1}$ and $ab \le a \log a - a + e^b$. More precisely, to get (22.74) from (22.71), one can write

$$N\rho \log Q = (N\rho^{1-\frac{1}{N}}) \rho^{\frac{1}{N}} \log Q \le (N\rho^{1-\frac{1}{N}}) \Big(\rho^{\frac{1}{N}} \log \rho^{\frac{1}{N}} - \rho^{\frac{1}{N}} + Q\Big);$$

and to get (22.75) from (22.71), one can write

$$N\rho^{1-\frac{1}{N}} \log \rho^{\frac{1}{N}} = (N\rho^{1-\frac{1}{N}} e^{-Q})(e^Q \log \rho^{\frac{1}{N}}) \le (N\rho^{1-\frac{1}{N}} e^{-Q}) \Big(e^Q Q - 2e^Q + e\rho^{\frac{1}{N}}\Big).$$

□

All the inequalities appearing in Corollary 22.36 can be seen as refinements of the Talagrand inequality appearing in Theorem 22.10; concentration inequalities derived from them take into account, for instance, the fact that the distance between any two points can never exceed $\pi(N-1)/\sqrt{K}$.

Exercise 22.37. Recover inequality (22.75) more directly by using the fact that $U(r) = r \log r$ lies in $\mathcal{DC}_N$.

Exercise 22.38. Use the inequalities proven in this section, and the result of Exercise 22.18, to recover, at least formally, the inequality

$$\left[\int h \, d\nu = 0\right] \Longrightarrow \qquad \|h\|^2_{H^{-1}(\nu)} \le \frac{KN}{N-1} \|h\|^2_{L^2(\nu)}$$

under an assumption of curvature-dimension bound $\mathrm{CD}(K,N)$. Now turn this into a rigorous proof, assuming as much smoothness on h and on the density of ν as you wish. (Hint: When $\varepsilon \to 0$, the optimal transport between $(1+\varepsilon h)\,\nu$ and ν converges in measure to the identity map; this enables to pass to the limit in the distortion coefficients.)

Remark 22.39. If one applies the same procedure to (22.73), one recovers a constant $K(Np)/(Np-1)$, which reduces to the correct constant only in the limit $p \to 1$. As for inequality (22.74), it leads to just K (which would be the limit $p \to \infty$).

Remark 22.40. Since the Talagrand inequality implies a Poincaré inequality without any loss in the constants, and the optimal constant in the Poincaré inequality is $KN/(N-1)$, it is natural to ask whether this is also the optimal constant in the Talagrand inequality. The answer is affirmative, in view of Theorem 22.15, since the logarithmic Sobolev inequality also holds true with the same constant. But I don't know of any transport proof of this fact!

Open Problem 22.41. *Find a direct transport argument to prove that the curvature-dimension* $\mathrm{CD}(K,N)$ *with* $K > 0$ *and* $N < \infty$ *implies* $T_2(\widetilde{K})$ *with* $\widetilde{K} = KN/(N-1)$, *rather than just* $T_2(K)$.

Note that inequality (22.75) does not solve this problem, since by Remark 22.39 it only implies the Poincaré inequality with constant K.

I shall conclude with a very loosely formulated open problem, which might be nonsense:

Open Problem 22.42. *In the Euclidean case, is there a particular variant of the Talagrand inequality which takes advantage of the homogeneity under dilations, just as the usual Sobolev inequality in* $\mathbb{R}^n$*? Is it useful?*

Bibliographical Notes

Most of the literature described below is reviewed with much more details in the synthesis works of Ledoux [237, 233, 234]. Selected applications to various parts of mathematics (Banach space theory, fine study of Brownian motion, combinatorics, percolation, spin glass systems, random matrices, etc.) are briefly developed in [237, Chapters 3 and 8]. The role of T_p inequalities in concentration theory is discussed in [237, Chapter 6], [21, Chapter 8], and [188]. One may also take a look at Massart's Saint-Flour lectures notes [258].

Lévy is often quoted as the founding father of concentration theory. His work might have been forgotten without the obstination of Milman to make it known. The modern period of concentration of measure starts with a work by Milman himself on the so-called Dvoretzy theorem [274].

The Lévy–Gromov isoperimetric inequality [193] is a way to get rather sharp concentration estimates from Ricci curvature bounds. Gromov has further worked on the links between Ricci curvature and concentration, see his very influential book [196], especially Chapter $3\frac{1}{2}$ therein. Also Talagrand made decisive contributions to the theory of concentration of measure, especially in product spaces, see in particular [346, 347].

Proposition 22.2 was studied by Rachev [307] and Bobkov and Götze [57], in the cases $p = 1$ and $p = 2$. These duality formulas were later systematically exploited by Ledoux and collaborators [59, 56]. The Legendre reformulation of the H functional can be found in many sources (for instance [247, Appendix B] when $\mathcal{X}$ is compact).

Proposition 22.5 goes back to Marton [257] and has been since then adapted to many situations, such as weakly dependent Markov chains; see [141] and the many references therein. The measurable selection theorem used in the construction of the coupling π can be found e.g. in [133]. As for Lemma 22.8, it is as old as information theory, since Shannon [**?**] used it to motivate the introduction of entropy in this context.

Marton [257] introduced the simple argument by which T_p inequalities lead to concentration inequalities (implication (i) $\Rightarrow$ (iii) in Theorem 22.9), and which has since then been reproduced in nearly all introductions to the subject. She used it mainly with the so-called *Hamming distance*: $d((x_i)_{1\le i\le n}, (y_i)_{1\le i\le n}) = \sum 1_{x_i \neq y_i}$.

Theorem 22.9 has been obtained by patching together results due to Bobkov and Götze [57], Djellout, Guillin and Wu [141], and Bolley and myself [63], together with a few arguments from folklore.

Weighted Csiszár–Kullback–Pinsker inequalities such as (22.13) were introduced in my paper with Bolley [63]; then Gozlan and Léonard [189] studied similar inequalities from the point of view of the theory of large deviations. More information can be found in Gozlan's PhD Thesis [188].

Relations of T_p inequalities with the so-called slicing problem are briefly discussed in [271].

Talagrand [348] proved Theorem 22.10 when ν is the Gaussian measure in $\mathbb{R}^n$, using a change of variables in the one-dimensional case, and then a tensorization argument.

This strategy was developed by Blower [53] who proved Theorem 22.10 when $M = \mathbb{R}^n$, $\nu(dx) = e^{-V(x)}\, dx$, $\nabla^2 V \geq K > 0$; more recently, it was used by Barthe [**?**] to recover the modified transport inequalities for the exponential measure on the half-line (a particular case of Theorem 22.21).

Otto and myself [292] found an alternative approach to Theorem 22.10, via the HWI inequality (which at that time of [292] had been established only in $\mathbb{R}^n$). The proof which I have used in this chapter is the same as the proof in [292], modulo the extension of the HWI inequality to general Riemannian manifolds.

There are several other schemes of proof for Theorem 22.10. One consists in combining Theorems 21.2 and 22.15. When $M = \mathbb{R}^n$, there is an argument based on Caffarelli's log concave perturbation theorem [91] (exercise). Yet another proof has been given by Bobkov and Ledoux [59], based on the Brunn–Minkowski theorem, or its functional counterpart the Prékopa–Leindler inequality (in this work there are interesting extensions to cases where the convexity assumptions are not the standard ones). Bobkov and Ledoux only worked in $\mathbb{R}^n$, but it is quite possible that their strategy can be extended to genuinely Riemannian situations, by means of the "Riemannian" Prékopa–Leindler inequality stated in Theorem 19.12.

Theorem 22.15 (log Sobolev implies T_2 implies Poincaré) was first proven by Otto and myself [292]; the Otto calculus had first been used to get an intuition of the result. Our proof relied on a heat semigroup argument, which will be explained later in Chapter 25. The "dual" strategy which I have used in this chapter, based on the Hamilton–Jacobi semigroup is due to Bobkov, Gentil and Ledoux [56]. In [292] it was assumed that the Ricci curvature of the manifold M is bounded below, and this assumption was removed in [56]. This is because the proof in [292] used a heat semigroup, which has infinite speed of propagation and is influenced by the asymptotic behavior of the manifold, while the argument in [56] was based on the Hopf–Lax semigroup, for which there is only finite speed of propagation (if the initial datum is bounded).

Various generalizations of the proof in [292] were considered by Cattiaux and Guillin [106].

The proof of [56] was adapted by Lott and myself [**?**] to compact length spaces $(\mathcal{X}, d)$ equipped with a reference measure ν that is locally doubling and satisfies a local Poincaré inequality; see Theorem 30.24 in the last chapter of these notes. In fact the proof of Theorem 22.15, as I have written it, is essentially a copy-paste from [247].

It was shown in [292] that (Talagrand) $\Rightarrow$ (log Sobolev) in $\mathbb{R}^n$, if the reference measure ν is log concave (with respect to the Lebesgue measure). It was natural to conjecture that the same argument would work under an assumption of nonnegative curvature (say $\mathrm{CD}(0, \infty)$); Theorem 22.19 shows that such is indeed the case.

Theorem 22.9 shows that T_1 is quite well understood, but such is not the case for T_2. It was only recently that Cattiaux and Guillin [106] produced a counterexample on the real line, showing that the T_2 inequality does not necessarily imply a log Sobolev inequality. Their counterexample takes the form $d\nu = e^{-V}\, dx$, where V oscillates rather wildly at infinity, in particular V'' is not bounded below. Counterexamples with V'' bounded below have still not yet been found.

Even more recently, Gozlan [**?**] exhibited a characterization of T_2 on $\mathbb{R}$, for certain classes of measures. This is still an active area of research.

The perturbation formula (22.23) for T_2 was first established by Blower [53] and later recovered with simpler methods by Bolley and myself [63].

There are alternative functional approaches to the concentration of measure: directly via logarithmic Sobolev inequalities [237, Chapter 5] [21, Chapter 7]; and via Brunn–Minkowski, Prékopa–Leindler, or isoperimetric inequalities [237, Chapter 2]. For instance,

(19.28) immediately implies

$$\nu[A^r] \geq 1 - \frac{e^{-\frac{Kr^2}{4}}}{\mu[A]}.$$

This kind of inequalities goes back to Gromov and Milman [197], who also were the first to study concentration from Poincaré inequalities. The tight links between all these functional inequalities show that these various strategies are in some sense related. First introduced by Herbst, the Laplace transform became an important tool in some of these developments, especially in the hands of Ledoux (see for instance [237]).

Theorem 22.10 admits an almost obvious generalization: if $\mathcal{F}$ is uniformly K-displacement convex and minimum at ν, then

$$\frac{K\, W_2(\mu,\nu)^2}{2} \leq \mathcal{F}(\mu) - \mathcal{F}(\nu). \tag{22.76}$$

Such inequalities have been studied in [292, 103, 15] and proven useful in the study of certain partial differential equations: see e.g. [103]. In Section 5 of this work, (22.76) is combined with the HWI inequality and the convergence of the functional $\mathcal{F}$, to deduce convergence in total variation. By the way, this is one of the rare instances that I know where the T_2 inequality has a real advantage on the T_1 inequality (apart from tensorization issues).

Optimal transport inequalities in *infinite dimension* have started to receive a lot of attention recently, for instance on the Wiener space. A major technical difficulty is that the natural distance in this problem, the so-called Cameron–Martin distance, takes value $+\infty$ "most of the time". Gentil [**?**] established the T_2 inequality for the Wiener measure by using the logarithmic Sobolev inequality on the Wiener space, and adapting the arguments of Bobkov, Gentil and Ledoux [56] based on Hamilton–Jacobi semigroup. Feyel and Üstünel [**?**] on one hand, Djellout, Guillin and Wu [141, Section 6] on the other hand, suggested a more direct approach based on Girsanov's formula. Interestingly enough, the T_2 inequality on the Wiener space implies the T_2 inequality on the Gaussian space, just by "projection" under the map $(x_t)_{0\leq t\leq 1} \to x_1$; this gives another proof of Talagrand's original inequality for the Gaussian measure. Wang [**?**] studied another kind of Talagrand inequality on the path space over an arbitrary Riemannian manifold.

In his recent PhD Thesis, Shao [**?**] studied T_2 inequalities on the path space and loop space constructed over a compact Lie group G. (The path space is equipped with the Wiener measure over G.) Together with Fang [**?**], he adapted the strategy based on the Girsanov formula, to get a T_2 inequality on the path space, and also on the path space over the loop space; then by reduction he gets a T_2 inequality on the loop space (equipped with a measure associated with the Brownian motion on loop space). This approach however only seems to give results when the loop space is equipped with the topology of uniform convergence, not with the more natural Cameron–Martin distance. I refer to [**?**] for more explanations.

Shao and Fang also extended Theorem 22.15 (Logarithmic Sobolev implies Talagrand inequality) to an infinite-dimensional setting, via the study of the Hamilton–Jacobi semigroup in infinite dimension. Thanks to known results about logarithmic Sobolev inequalities on loop spaces (studied by Driver, Lohrentz and others), they recover a T_2 inequality on the loop space, now for the Cameron–Martin distance. The technical core of these results is the analysis of the Hamilton–Jacobi for semi-distances in infinite dimension, performed in [**?**].

Very recently, Shao and Fang [**?**] used Talagrand inequalities to obtain results of unique existence of optimal transport in the Wiener space over a Lie group, when the target

measure ν is the Wiener measure and the source measure μ satisfies $H_\nu(\mu) < +\infty$. In the standard (Gaussian) Wiener space, Feyel and Üstünel have solved the same problem in more generality, but so far their results have not been extended outside the Gaussian setting.

The equivalence between Poincaré inequalities and modified transport inequalities, expressed in Theorem 22.21, has a long history. Talagrand [**?**] had identified concentration properties satisfied by the exponential measure, or a product of exponential measures. He showed the following precised version of (22.69):

$$\nu^{\otimes N}\big[A + 6\sqrt{r}B_1^{d_2} + 9rB_1^{d_1}\big] \geq 1 - \frac{e^{-r}}{\nu^{\otimes N}[A]}.$$

A proof can be found in [237, Theorem 4.16]. It is also Talagrand who noticed that concentration inequalities for the product exponential measure were in some sense stronger than concentration inequalities for the Gaussian measure (Remark 22.31 and Example 22.33, which I copied from [237]). Then Maurey [**?**] found a simple approach to concentration inequalities for the product exponential measure. Later Talagrand [348] made the connection with transport inequalities with quadratic-linear cost. Bobkov and Ledoux [**?**] introduced modified logarithmic Sobolev inequalities, and showed their equivalence with Poincaré inequalities. The proof of (i) $\Rightarrow$ (ii) is copied almost verbatim from [**?**]. Bobkov and Ledoux also showed how to recover concentration inequalities directly from these modified logarithmic Sobolev inequalities, showing in some sense that the concentration properties of the exponential measure were shared by all measures satisfying a Poincaré inequality. Finally, Bobkov, Gentil and Ledoux [56] understood how to deduce quadratic-linear transportation inequalities from modified logarithmic Sobolev inequalities, thanks to the Hamilton–Jacobi semigroup. The proof of Theorem 22.25 is just an expanded version of the arguments suggested in [56].

In the particular case when $\nu(dx) = e^{-|x|}\,dx$ on $\mathbb{R}_+$, there are simpler proofs of Theorem **??**, also with improved constants.

The treatment of dimension-dependent Talagrand-type inequalities in the last section is inspired from a joint work with Lott [249]. That topic had been addressed before, with different tools, by Gentil [182]; it would be interesting to compare precisely his results with the ones in this chapter.

23

Gradient flows I: Definition and convergence

Take a Riemannian manifold M and a function $\Phi : M \to \mathbb{R}$, which for the sake of this exposition will be assumed to be continuously differentiable. The gradient of Φ, denoted by $\nabla\Phi$, is the vector field defined by the equation

$$d_x\Phi \cdot v = \langle \nabla_x \Phi, v \rangle_x,$$

where v is an arbitrary vector in the tangent space T_xM, $d_x\Phi$ stands for the differential of Φ at x, and $\langle \cdot, \cdot \rangle_x$ is the scalar product on T_xM. In other words, if $(\gamma_t)_{-\varepsilon<t<\varepsilon}$ is a smooth path in M, with $\gamma_0 = x$, then

$$\left[\left.\frac{d}{dt}\right|_{t=0} x_t = v \right] \Longrightarrow \qquad \left.\frac{d}{dt}\right|_{t=0} \Phi(\gamma_t) = \langle \nabla_x \Phi, v \rangle_x.$$

If $|v|$ is given, then in order to make the latter derivative as large as possible, the best choice is to take v colinear to $\nabla_x\Phi$. In that sense $\nabla_x\Phi$ indicates *the direction in which Φ increases most rapidly.*

Now the gradient flow associated to Φ is the flow defined by the differential equation

$$\frac{dX}{dt} = -\operatorname{grad}_X \Phi.$$

One may think of it heuristically as a flow which tries to *make Φ decrease as fast as possible.* Stated in this way, this intuition is of course grossly false: for instance, $\dot{X} = -\lambda \operatorname{grad}_X \Phi$, $\lambda > 1$, will make Φ decrease even faster; but later in this chapter I shall make the statement more precise and more convincing.

An important consequence of the definition of gradient flow is the following neat formula for the time-derivative of the energy:

$$\frac{d}{dt}\Phi(X(t)) = -\left|\operatorname{grad}_{X(t)}\Phi\right|^2.$$

Gradient flows (as Hamiltonian flows) are everywhere in physics and mathematics. In mechanics, they often describe the behavior of Hamiltonian systems, in an asymptotic regime in which dissipative effects play such an important role, that the effects of forcing and dissipation compensate each other. The basic example one should think of is

$$\ddot{X} = -\lambda \operatorname{grad}_X \Phi - \lambda \dot{X}$$

(acceleration = forcing - friction), in the limit $\lambda \to \infty$ (strong friction).

Gradient flows in Wasserstein space

Around the end of the nineties, Jordan, Kinderlehrer and Otto made the important discovery that *many important partial differential equations can be reformulated as gradient flows in the Wasserstein space*. The most emblematic example is that of the heat equation,

$$\partial_t \rho = \Delta \rho,$$

say in Euclidean space for simplicity. It was well-known that this equation can be seen as a gradient flow, for instance for the quadratic functional $\Phi(\rho) = \int |\nabla \rho|^2 \, dx$ in $L^2(\mathbb{R}^n)$. But the Jordan–Kinderlehrer–Otto formulation describes the heat equation as a gradient flow *in the space of probability measures*, with a natural "information-theoretical" content. In this new approach, the functional Φ is the Kullback information, or negative of the entropy, $\Phi(\rho) = \int \rho \log \rho \, dx$.

To better understand this point of view, Otto developed what I dubbed "Otto calculus" in Chapter 15. We have already saw several applications of this calculus, at least for heuristic purposes.

In this chapter, I shall describe in which *rigorous* sense one can say that certain equations are gradient flows in the Wasserstein space. Before that, it will be necessary to get a good understanding of gradient flows in abstract metric spaces, a subject which is important in itself.

Reformulations of gradient flows

There are several ways to reformulate gradient flows in a weak sense, so as to obtain definitions that are general (for nonsmooth energies, or nonsmooth spaces), and stable (under some limit process). They usually require a convexity-type assumption on the energy Φ. Here I shall present some of these reformulations and explain why they are equivalent to the classical formulation when used in a smooth setting.

Proposition 23.1 (reformulations of gradient flows). *Let M be a Riemannian manifold, let $\Lambda = \Lambda(x, v)$ be a quadratic form on TM, bounded below, and let Φ be a C^1 function $M \to \mathbb{R}$, Λ-convex in the sense of Proposition 16.2. Let further $X : (t_1, t_2) \to M$, and let $t \in (t_1, t_2)$ be a time where X is differentiable. Then, the following statements are equivalent:*

(i) $\dot{X}(t) = -\operatorname{grad}_{X(t)} \Phi$;

(ii) $-\dot{X}(t) \in \partial \Phi(X(t))$;

(iii) $\dfrac{|\dot{X}(t)|^2 + |\nabla^- \Phi(X(t))|^2}{2} = -\dfrac{d}{dt} \Phi(X(t))$;

(iv) for any $y \in M$, and for any geodesic $(\gamma_s)_{0 \le s \le 1}$ with $\gamma_0 = y$, $\gamma_1 = X(t)$,

$$\frac{d^+}{dt} \left(\frac{d(y, X(t))^2}{2} \right) \le \Phi(y) - \Phi(X(t)) - \int_0^1 \Lambda(\gamma_s, \dot{\gamma}_s) \, s \, ds.$$

Remark 23.2. The most well-known case is when $\lambda = 0$ (Φ is convex), and then (iv) becomes just

$$\frac{d^+}{dt} \left(\frac{d(y, X(t))^2}{2} \right) \le \Phi(y) - \Phi(X(t)).$$

Remark 23.3. Statements (i) to (iii) do not explicitly depend on Λ, so here the assumption of Λ-convexity is not essential. But as soon as one wants to generalize Proposition 23.1 by dropping some smoothness assumptions, it might be important to know that Φ is Λ-convex for some Λ. Note that in formulation (iv), one can always replace Λ by $\Lambda' \le \Lambda$, and the equivalence still holds true, independently of the choice of Λ'! Accordingly, in practice one can restrict to the choice $\Lambda(x, v) = \lambda|v|^2$, that is, when Φ is λ-convex.

Remark 23.4. If one wants to use Proposition 23.1 to characterize a curve $(X(t))$ as a gradient flow, the natural regularity assumption is that X be an absolutely continuous function of t. This will imply the existence of the derivative $\dot{X}(t)$ for almost all t, and in addition this will guarantee that the values of X are uniquely determined by $X(0)$ and the values of $\dot{X}$.

Before going on with the proof of Proposition 23.1, I shall briefly explain its interest. Property (ii) has the advantage to be formulated in terms of subdifferentials, which are well-defined for semi-convex functions (smooth or not), and quite stable. Property (iii) involves speeds (norms of velocities) rather than velocities; this is interesting also in a nonsmooth setting, where the speed might be well-defined even if the velocity is not. Finally, Property (iv) appears to be quite handy to study gradient flows in an abstract metric space; this is the one that I shall use in the sequel (with the particular choice $\Lambda(x, \cdot) = \lambda|\cdot|^2$, which is no loss of generality as explained in Remark 23.3).

Proof of Proposition 23.1. First recall some definitions:

$$|\nabla^- \Phi(x)| = \limsup_{y \to x} \frac{[\Phi(y) - \Phi(x)]_-}{d(x,y)};$$

$$\partial\Phi(x) = \Big\{ v \in T_x M; \quad \forall w \in T_x M,\ \Phi\big(\exp_x(\varepsilon w)\big) \ge \Phi(x) + \varepsilon\langle v, w\rangle + o(\varepsilon) \Big\}.$$

Since Φ is differentiable by assumption,

$$|\nabla^- \Phi(x)| = |\nabla\Phi(x)|, \qquad \partial\Phi(x) = \{\nabla\Phi(x)\}.$$

Then the equivalence between (i) and (ii) is obvious.

Next, by chain-rule, Cauchy–Schwarz and Young's inequalities,

$$-\frac{d}{dt}\Phi(X(t)) = \Big\langle -\nabla\Phi(X(t)),\, \dot{X}(t) \Big\rangle \le \big|\nabla\Phi(X(t))\big|\,\big|\dot{X}(t)\big| \le \frac{|\nabla\Phi(X(t))|^2 + |\dot{X}(t)|^2}{2},$$

with equality if and only if $-\nabla\Phi(X(t))$ and $\dot{X}(t)$ have the same norm and opposite directions. So (i) is equivalent to (iii).

Now, let us check the equivalence of (i) and (iv). Let y be given, and let $\gamma(s)$ be a geodesic path joining $\gamma(0) = y$ to $\gamma(1) = X(t)$. Then by the formula of first variation (as in the Appendix of Chapter 7),

$$\frac{d^+}{dt}\left(\frac{d(y, X(t))^2}{2}\right) \le \big\langle \dot{\gamma}(1), \dot{X}(t) \big\rangle_{X(t)} \tag{23.1}$$

(the distance increases if $\dot{X}$ is in the direction of $\dot{\gamma}(1)$). On the other hand, since Φ is Λ-convex,

$$\Phi(\gamma(0)) \ge \Phi(\gamma(1)) - \Big\langle \dot{\gamma}(1),\, \nabla\Phi(\gamma(1)) \Big\rangle + \int_0^1 \Lambda\big(\gamma(s), \dot{\gamma}(s)\big)\, s\, ds.$$

(This by the way does not depend on Φ being C^1.) So

$$\Big\langle \dot\gamma(1),\, -\nabla\Phi(X(t))\Big\rangle \le \Phi(y) - \Phi(X(t)) - \int_0^1 \Lambda\big(\gamma(s),\dot\gamma(s)\big)\, s\, ds.$$

This combined with (23.1) proves the implication (i) $\Rightarrow$ (iv).

For the reverse implication, let again t be given, $w \in T_{X(t)}M$, $y = \exp_{X(t)}(\varepsilon w)$. If ε is small enough there is a unique geodesic γ joining $\gamma(0) = X(t)$ to $\gamma(1) = y$, namely $\gamma(s) = \exp_{X(t)}(s\varepsilon w)$. Then $|\dot\gamma| = \varepsilon|w|$, and

$$\frac{d}{dt}\left(\frac{d(y,X(t))^2}{2}\right) = \big\langle \dot\gamma(1), \dot X(t)\big\rangle = -\big\langle \varepsilon w, \dot X(t)\big\rangle.$$

So if Property (iv) is satisfied, then

$$\begin{aligned}
\big\langle \varepsilon w, -\dot X(t)\big\rangle &= \frac{d^+}{dt}\left(\frac{d(y,X(t))^2}{2}\right)\\
&\le \Phi(y) - \Phi(X(t)) - \int_0^1 \Lambda\big(\gamma(s),\dot\gamma(s)\big)\, s\, ds\\
&\le \Phi(\exp_{X(t)}\varepsilon w) - \Phi(X(t)) + \lambda\int_0^1 |\dot\gamma(s)|^2\, s\, ds\\
&\le \Phi(\exp_{X(t)}\varepsilon w) - \Phi(X(t)) + \lambda\,\frac{\varepsilon^2}{2}.
\end{aligned}$$

As a consequence,

$$\Phi(\exp_{X(t)}\varepsilon w) \ge \Phi(X(t)) + \varepsilon\big\langle w, -\dot X(t)\big\rangle + o(\varepsilon),$$

which precisely means that $-\dot X(t) \in \partial\Phi(X(t))$, so (ii) is satisfied. This shows that (iv) implies (ii).

In the end, all Properties (i)-(iv) are equivalent, which proves Proposition 23.1. □

Gradient flows in metric spaces

Proposition 23.1 suggests the following definition as a possible way to introduce gradient flows in possibly nonsmooth length spaces.

Definition 23.5 (Gradient flows in length space). *Let $(\mathcal{X}, d)$ be a length space, $\lambda \in \mathbb{R}$, and let Φ be a lower semi-continuous, λ-convex functional $\mathcal{X} \to \mathbb{R}\cup\{+\infty\}$. Let $(X(t))_{t\ge 0}$ be a path in $\mathcal{X}$, absolutely continuous in the sense of* (7.5)*. Then X is said to be a trajectory of the gradient flow associated with the energy Φ if, for all $y \in \mathcal{X}$ and all $t > 0$,*

$$\frac{d^+}{dt}\left(\frac{d(y,X(t))^2}{2}\right) \le \Phi(y) - \Phi(X(t)) - \lambda\,\frac{d(y,X(t))^2}{2}.$$

Proposition 23.1 guarantees that this concept coincides with the usual one when $\mathcal{X}$ is a Riemannian manifold equipped with its geodesic distance. The following statement shows that gradient flows are relatively well-behaved; it also provides a simple sufficient condition for a trajectory to be a gradient flow.

Proposition 23.6 (Properties of gradient flows). *Let $(\mathcal{X}, d)$, λ and Φ be as in Definition 23.5, and let $(X(t))_{t\geq 0}$ be a path in $\mathcal{X}$, absolutely continuous. Then*

(i) If one has, for almost all $t > 0$,

$$\frac{d}{dt}\left(\frac{d(y, X(t))^2}{2}\right) \leq \left.\frac{d^+}{ds}\right|_{s=0} \Phi(\gamma(s)),$$

where $\gamma(s)$ is a (constant-speed, minimizing) geodesic joining $\gamma(0) = X(t)$ to $\gamma(1) = y$, then X is a trajectory of the gradient flow associated with Φ;

(ii) If X and $\widetilde{X}$ are two trajectories of the gradient flow associated with Φ, then

$$\forall t \geq 0, \qquad d\big(X(t), \widetilde{X}(t)\big) \leq e^{\lambda t} d\big(X(0), \widetilde{X}(0)\big).$$

In particular, there can be at most one trajectory of the gradient flow starting from a given initial condition.

Remark 23.7. Property (ii) guarantees that Definition 23.5 does not depend on the choice of λ.

Proof of Proposition 23.6. First, it is a general fact that if Φ is λ-convex in a length space, and γ is a geodesic joining $\gamma(0) = X(t)$ to $\gamma(1) = y$, then

$$\left.\frac{d^+}{ds}\right|_{s=0} \Phi(\gamma(s)) \leq \Phi(y) - \Phi(X(t)) - \lambda\, \frac{d(y, X(t))^2}{2}.$$

This implies (i) at once.

To prove (ii), fix $T > 0$, and define

$$F(s,t) := \frac{d\big(X(s), \widetilde{X}(t)\big)^2}{2}, \qquad 0 \leq s, t \leq T.$$

If C is a bound for $d(X(s), \widetilde{X}(t))$, then

$$\begin{aligned}\big|F(s,t) - F(s',t)\big| &\leq \left(\frac{d(X(s), \widetilde{X}(t)) + d(X(s'), \widetilde{X}(t))}{2}\right) \Big|d\big(X(s), \widetilde{X}(t)\big) - d\big(X(s'), \widetilde{X}(t)\big)\Big| \\ &\leq C\, d\big(X(s), X'(s)\big).\end{aligned}$$

Since X is an absolutely continuous path, it follows that F is *uniformly* (in t) absolutely continuous with respect to s. By Lemma 23.19 in the Appendix,

$$\begin{aligned}\left.\frac{d}{dt}\right|_{t=t_0} F(t,t) &= \left.\frac{d}{dt}\right|_{t=t_0} \frac{d\big(X(t_0), \widetilde{X}(t)\big)^2}{2} + \left.\frac{d}{dt}\right|_{t=t_0} \frac{d\big(\widetilde{X}(t_0), X(t)\big)^2}{2} \\ &\leq \left[\Phi(X(t_0)) - \Phi(\widetilde{X}(t_0)) - \lambda\, \frac{d(X(t_0), \widetilde{X}(t_0))^2}{2}\right] \\ &\qquad + \left[\Phi(\widetilde{X}(t_0)) - \Phi(X(t_0)) - \lambda\, \frac{d(\widetilde{X}(t_0), X(t_0))^2}{2}\right] \\ &= -\lambda\, d(X(t_0), \widetilde{X}(t_0))^2 \\ &= -2\lambda F(t_0, t_0).\end{aligned}$$

It follows that $F(t,t) \leq e^{-2\lambda t} F(0,0)$, whence the conclusion. □

In the sequel, I shall apply Definition 23.5 in the Wasserstein space $\mathcal{X} = P_2(M)$, where M is a smooth Riemannian manifold (sometimes with additional assumptions of bounds on the Hessians of the square distance). To avoid complications I shall use Definition 23.5 in $P_2^{\rm ac}(M)$, that is, restricting to absolutely continuous probability measures. This might look a bit dangerous, because $P_2^{\rm ac}(M)$ is not complete, but after all it is a length space on its own right, as a geodesically convex subset of $P_2(M)$ (Recall Theorem 8.5(ii)), and I shall not need completeness. Of course, this does not mean that it is not interesting to study gradient flows in the whole of $P_2(M)$.

To go on with this program, I should first

- compute the (upper) derivative of the distance function;

- compute the subdifferential of a given energy functional.

This will be the object of the next two sections.

Derivative of the Wasserstein distance

Theorem 23.8 (Derivative of the Wasserstein distance). *Let M be a smooth Riemannian manifold. Let (μ_t) and $(\widetilde{\mu}_t)$ be two curves $(t_1, t_2) \to P_2^{\rm ac}(M)$, weak solution of the continuity equations*

$$\frac{\partial \mu_t}{\partial t} + \nabla \cdot (\xi_t\, \mu_t) = 0, \qquad \frac{\partial \widetilde{\mu}_t}{\partial t} + \nabla \cdot (\widetilde{\xi}_t\, \widetilde{\mu}_t) = 0, \tag{23.2}$$

where $\xi_t = \xi_t(x)$, $\widetilde{\xi}_t = \widetilde{\xi}_t(x)$ are uniformly Lipschitz functions of t and x. Then $t \to \mu_t$ and $t \to \widetilde{\mu}_t$ are absolutely continuous curves. Moreover, if $t \in (t_1, t_2)$ is given, then

$$\frac{d}{dt}\left(\frac{W_2(\sigma, \mu_t)^2}{2}\right) = -\int_M \langle \widetilde{\nabla}\psi, \xi\rangle\, d\mu_t \;-\; \int_M \langle \widetilde{\nabla}\widetilde{\psi}, \widetilde{\xi}\rangle\, d\widetilde{\mu}_t,$$

where ψ, $\widetilde{\psi}$ are $(d^2/2)$-functions such that

$$\exp(\widetilde{\nabla}\psi)_\# \mu_t = \widetilde{\mu}_t, \qquad \exp(\widetilde{\nabla}\widetilde{\psi})_\# \widetilde{\mu}_t = \mu_t.$$

Remark 23.9. Recall that Theorem 10.35 gives a list of a few conditions under which the approximate gradient $\widetilde{\nabla}$ can be replaced by the usual gradient ∇ in the formulas above.

Remark 23.10. For the purpose of this chapter, the superdifferentiability of the Wasserstein distance would be enough. However, for the sake of completeness, I shall also establish the subdifferentiability, which is a bit more tricky.

Proof of Theorem 23.8. Let us first consider the case when the path $\widetilde{\mu}_t$ is constant and equal to some fixed measure $\sigma \in P_2^{\rm ac}(M)$.

Step 1: Superdifferentiability. Let T be the optimal (Monge) transport $\sigma \to \mu_t$. Then

$$\frac{W_2(\sigma, \mu_t)^2}{2} = \frac{1}{2}\int d(x, T(x))^2\, d\sigma(x). \tag{23.3}$$

For any $\tau > 0$, we can construct a family of trajectories on $[t, t+\tau]$ by following the velocity field ξ:

$$\frac{d}{ds}\Big(T_{t\to t+s}(x)\Big) = \xi_{t+s}\big(T_{t\to t+s}(x)\big).$$

The maps $T_{t\to t+s}$ are well-defined and locally Lipschitz, as a consequence of the Cauchy–Lipschitz theory. Moreover, by the usual Eulerian/Lagrangian duality and the continuity equation (23.2),

$$(T_{t\to t+s})_\# \mu_t = \mu_{t+s}.$$

So $T_{t\to t+s} \circ T$ is a transport $\sigma \to \mu_{t+s}$, and by definition of the Wasserstein distance,

$$\frac{W_2(\sigma, \mu_{t+s})^2}{2} \le \frac{1}{2} \int d\Big(x, T_{t\to t+s} \circ T(x)\Big)^2 \, d\sigma(x).$$

This, combined with (23.3), implies

$$\frac{1}{s}\left(\frac{W_2(\sigma, \mu_{t+s})^2}{2} - \frac{W_2(\sigma, \mu_t)^2}{2}\right) \le \int \left(\frac{d\big(x, T_{t\to t+s} \circ T(x)\big)^2 - d(x, T(x))^2}{2s}\right) d\sigma(x). \tag{23.4}$$

Since there is a minimizing geodesic connecting $T(x)$ to x with initial velocity $\widetilde{\nabla}\psi(T(x))$, the formula of first variation yields

$$\forall x \in M, \qquad \limsup_{s\downarrow 0} \left[\frac{d\big(x, T_{t\to t+s} \circ T(x)\big)^2 - d(x, T(x))^2}{2s}\right] \le -\big\langle \xi_t(T(x)), \widetilde{\nabla}\psi(T(x))\big\rangle.$$

So if we can pass to the lim sup as $s \to 0$ in (23.4), it will follow that

$$\begin{aligned}
\frac{d^+}{dt}\left(\frac{W_2(\sigma, \mu_t)^2}{2}\right) &\le -\int_M \big\langle \xi_t(T(x)), \widetilde{\nabla}\psi(T(x))\big\rangle \, d\sigma(x) \\
&= -\int \langle \xi_t(y), \widetilde{\nabla}\psi(y)\rangle \, d(T_\#\sigma)(y) \\
&= -\int \big\langle \xi_t(y), \widetilde{\nabla}\psi(y)\big\rangle \, d\mu_t(y),
\end{aligned}$$

and this will establish the right-superdifferentiability of $W_2(\mu_t, \sigma)$.

So we should check that we can indeed pass to the lim sup in (23.4). For this we can use Fatou's lemma, in the following form: *If $v = v(s,x)$ is a real-valued function, bounded above by $w(x)$, where $w \in L^1(d\sigma)$, then*

$$\limsup_{s\to 0} \int v(s,x)\, d\sigma(x) \le \int \Big[\limsup_{s\to 0} v(s,x)\Big] \, d\sigma(x).$$

In the present case, because ξ is Lipschitz, it is not difficult to derive the following estimates:

$$d\big(y, T_{t\to t+s}(x)\big) \le C(\tau)\big(1 + d(y,x)\big); \qquad d\big(x, T_{t\to t+s}(x)\big) \le C(\tau)\, s\big(1 + d(y,x)\big),$$

where y is some arbitrary point in M, and $C(\tau)$ is a constant which might depend on the upper bound of the time-interval. In the sequel, I shall not recall the dependence on τ, and I shall use the same symbol C to denote various such constants. Then

$$v(s,x) := \frac{d\big(x, T_{t\to t+s}\circ T(x)\big)^2 - d(x,T(x))^2}{2s}$$

$$\begin{aligned}
&\le \left|\frac{d\big(x, T_{t\to t+s}\circ T(x)\big) - d(x,T(x))}{s}\right| \left(\frac{d\big(x, T_{t\to t+s}\circ T(x)\big) + d(x,T(x))}{2}\right)\\
&\le \frac{d\big(T(x), T_{t\to t+s}(T(x))\big)}{s}\left(d(x,T(x)) + \frac{d\big(T(x), T_{t\to t+s}(T(x))\big)}{2}\right)\\
&\le C\big(1+d(y,T(x))\big)\Big(1+d(x,T(x))+d(y,T(x))\Big)\\
&\le C\Big(1+d(y,T(x))^2+d(x,T(x))^2\Big) =: w(x).
\end{aligned}$$

Then w is σ-integrable, since

$$\int d(y,T(x))^2\,d\sigma(x) = \int d(y,x)^2\,d\mu_t(x) < +\infty;$$
$$\int d(x,T(x))^2\,d\mu_t(x) = W_2^2(\sigma,\mu_t)^2 < +\infty.$$

This concludes the proof of the right-superdifferentiability of $W_2(\mu_t,\sigma)$. The same reasoning can be repeated to prove the left-superdifferentiability (that is, when $s\uparrow 0$), and the results is the same; so $W_2(\mu_t,\sigma)$ is in fact superdifferentiable. (Equivalently, one may reverse time by replacing the vector field ξ by $-\xi$.)

Step 2: Subdifferentiability. The reader might skip the rest of the proof at first reading. To establish the subdifferentiability, it is sufficient to establish the right-subdifferentiability, that is, only consider

$$\liminf_{s\downarrow 0} \frac{W_2(\mu_{t+s},\sigma)^2 - W_2(\mu_t,\sigma)^2}{s};$$

then the true subdifferentiability will follow by an argument similar to the one used before.

For each $s\neq 0$, let $T^{(s)}$ be the optimal transport between σ and μ_s. As $s\downarrow 0$ we can extract a subsequence $s_k\to 0$, such that

$$\limsup_{s\downarrow 0} \frac{W_2(\mu_t,\sigma)^2 - W_2(\mu_{t+s},\sigma)^2}{s} = \lim_{k\to\infty} \frac{W_2(\mu_{t+s_k},\sigma)^2}{s_k}.$$

Then, by reasoning as in Step 1,

$$\limsup_{s\downarrow 0} \frac{W_2(\mu_t,\sigma)^2 - W_2(\mu_{t+s},\sigma)^2}{s} \le \limsup_{k\to\infty} \int \frac{d\Big(x, T_{s_k+t\to t}\circ T^{(t+s_k)}(x)\Big)^2 - d\big(x, T^{(t+s_k)}(x)\big)^2}{s_k}\,\sigma(dx). \tag{23.5}$$

In other words,

$$\limsup_{s\downarrow 0} \frac{W_2(\mu_t,\sigma)^2 - W_2(\mu_{t+s},\sigma)^2}{s} \le \int v_k(x)\,\sigma(dx), \tag{23.6}$$

where

$$v_k(x) = \frac{d\Big(x, T_{t+s_k\to t}\circ T^{(t+s_k)}(x)\Big)^2 - d\big(x, T^{(t+s_k)}(x)\big)^2}{s_k}.$$

Since $T^{(t)}$ is the unique optimal transport between σ and μ_t, and since $s \to \mu_{t+s}$ is continuous with respect to the weak topology, we know from Corollary 5.20 that $T^{(t+s_k)}$ converges to $T^{(t)}$ in probability, with respect to the measure σ. Extracting a further subsequence if necessary, we may assume that $T^{(t+s_k)}$ converges to $T^{(t)}$ almost surely.

Next, the square distance d^2 is locally superdifferentiable, so

$$\begin{aligned}\frac{d\big(x, T_{t+s_k\to t}(x)\big)^2}{2} &\le \frac{d(x,y)^2}{2} + s_k\,\langle \xi_t(y), \dot\gamma(1)\rangle_y + o\big(d(y, T_{t+s_k\to t}(y)\big)\\ &\le \frac{d(x,y)^2}{2} + s_k\,\langle \xi_t(y), \dot\gamma(1)\rangle_y + o(s_k),\end{aligned}$$

where γ is the geodesic joining x to y, and the $o(s_k)$ is uniform in a neighborhood of y. So if $y_k \to y$, then

$$\limsup_{k\to\infty} \frac{d\big(x, T_{t+s_k\to t}(y_k)\big)^2 - d(x,y_k)^2}{s_k} \le \langle \xi_t(y), \dot\gamma(1)\rangle.$$

Applying this to $y_k = T^{(t+s_k)}(x) \to T^{(t)}(x)$, we deduce that

$$\limsup_{k\to\infty} v_k(x) \le v(x) := \big\langle \xi_t(T^{(t)}(x)), \dot\gamma(1)\big\rangle_{T^{(t)}(x)}.$$

Here again γ is the geodesic joining x to $T^{(t)}(x)$, so $\dot\gamma(1) = -\widetilde\nabla\psi(T^{(t)}(x))$, where $\exp(\widetilde\nabla\psi)$ is the optimal transport going from μ_t to σ. So to complete the proof, it is sufficient to pass to the lim sup in (23.6).

To pass to the limit, we need some control on the functions v_k. Let z be a fixed point in M. Since ξ is uniformly Lipschitz, it is easy to establish the bound $d(y, T^{(t+s_k)}(y)) \le C\,s_k\,(1 + d(z,y))$, where C is a constant; and as a consequence

$$\begin{aligned}&d\Big(x, T_{s_k+t\to t}\circ T^{(t+s_k)}(x)\Big)^2 - d\big(x, T^{(t+s_k)}(x)\big)^2\\ &\quad\le \Big[2d(x, T^{(t+s_k)}(x)) + d\Big(T^{(t+s_k)}(x), T_{t+s_k\to t}((T^{(t+s_k)}(x)))\Big)\Big]\; d\Big(T^{(t+s_k)}(x), d\big(T_{t+s_k\to t}\circ T^{(t+s_k)}(x)\big)\Big)\\ &\le C\Big(d(x, T^{(t+s_k)}(x)) + s_k\Big)\; s_k\,(1 + d(z,x)).\end{aligned}$$

It follows that

$$v_k(x) \le C\,\Big[1 + d(x, T^{(t+s_k)}(x))\Big]\,(1 + d(z,x)) \tag{23.7}$$

Let χ be a cut-off continuous function, $0 \le \chi \le 1$, $\chi(d) = 1$ for $d \le 1$, $\chi(d) = 0$ for $d \ge 2$, and let $\chi_R(d) = \chi(d/R)$. (This is a continuous approximation of $1_{d\le R}$.) When $\chi_R(d(z,x) + d(z, T^{(t+s_k)}(x))) \ne 0$, $v_k(x)$ stays bounded. So we can invoke Fatou's lemma as in Step 1:

$$\limsup_{k\to\infty} \int \chi_R\big(d(x, T^{(t+s_k)}(x))\big)\, v_k(x)\,\sigma(dx) \le \int \chi_R\big(d(x, T^{(t)}(x))\big)\, v(x)\,\sigma(dx).$$

To conclude the argument it suffices to show that

$$\lim_{R\to\infty} \left| \int \chi_R\big(d(x, T^{(t)}(x))\big)\, v(x)\,\sigma(dx)\right| = 0; \tag{23.8}$$

$$\lim_{R\to\infty}\ \limsup_{k\to\infty} \left| \int \chi_R\big(d(x, T^{(t+s_k)}(x))\big)\, v_k(x)\,\sigma(dx)\right| = 0; \tag{23.9}$$

Taking into account the bound $|v(x)| \le d(x, T^{(t)}(x)) \le$, we have

$$\begin{aligned}
&\left| \int (1 - \chi_R)\big(d(x, T^{(t)}(x))\big)\, v(x)\, \sigma(dx) \right| \\
&\le \int_{d(x,T^{(t)}(x)) \ge R} d(x, T^{(t)}(x))\, \sigma(dx) \\
&\le \frac{1}{R} \int d(x, T^{(t)}(x))^2\, \sigma(dx) \\
&= \frac{1}{R} W_2(\sigma, \mu_t)^2;
\end{aligned}$$

this establishes (23.8).

Similarly,

$$\left| \int (1 - \chi_R)\big(d(x, T^{(t+s_k)}(x))\big)\, v_k(x)\, \sigma(dx) \right| \le \frac{1}{R} W_2(\sigma, \mu_{t+s_k})^2,$$

which proves (23.9). This concludes the proof of the subdifferentiability.

Step 3: Doubling variables. At this stage we know that $s \to W_2(\mu_s, \widetilde{\mu}_t)$ is differentiable, and of course $s \to W_2(\mu_t, \widetilde{\mu}_s)$ is differentiable too. To conclude to the differentiability of $t \to W_2(\mu_t, \widetilde{\mu}_t)$, we can use again Lemma 23.19 in the Appendix, provided that we check that, say, $s \to W_2(\mu_s, \widetilde{\mu}_t)$ is (locally) absolutely continuous in s, uniformly in t. To prove this, we can use the triangular inequality, in the form

$$\begin{aligned}
W_2(\mu_s, \widetilde{\mu}_t)^2 - W_2(\mu_{s'}, \widetilde{\mu}_t)^2 &= \Big[W_2(\mu_s, \widetilde{\mu}_t) + W_2(\mu_{s'}, \widetilde{\mu}_t)\Big]\, \Big[W_2(\mu_s, \widetilde{\mu}_t) - W_2(\mu_{s'}, \widetilde{\mu}_t)\Big] \\
&\le \Big[W_2(\mu_s, \widetilde{\mu}_t) + W_2(\mu_{s'}, \widetilde{\mu}_t)\Big]\, W_2(\mu_s, \mu_{s'}) \\
&\le \Big[W_2(\mu_s, \sigma) + W_2(\mu_{s'}, \sigma) + 2\, W_2(\widetilde{\mu}_t, \sigma)\Big]\, W_2(\mu_s, \mu_{s'}),
\end{aligned}$$

where σ is any arbitrary element of $P_2(M)$. The quantity inside square brackets is bounded (in fact it is a Lipschitz function of s, s' and t), and the path (μ_s) is Lipschitz in W_2 distance; so in fact

$$W_2(\mu_s, \widetilde{\mu}_t)^2 - W_2(\mu_{s'}, \widetilde{\mu}_t)^2 \le C\, |s - s'|$$

for some constant C. The conclusion follows. □

Subdifferential of energy functionals

The problem addressed in the present section is to estimate the derivative of an energy functional such as U_ν, along a path in the Wasserstein space $P_2(M)$, or rather in $P_2^{\mathrm{ac}}(M)$. This problem is easy to solve formally, but a rigorous justification is definitely not trivial, and for the moment the only known strategy uses some mildly advanced real analysis, in particular Alexandrov's second differentiability theorem (Theorem 14.1), and (at least in the noncompact case) the notion of approximate gradient $\widetilde{\nabla}$.

Theorem 23.11 (Computation of subdifferentials in Wasserstein space). *Let M be a Riemannian manifold, equipped with a reference measure $\nu = e^{-V}\mathrm{vol}$ satisfying the curvature-dimension bound $\mathrm{CD}(K, N)$ for some $K \in \mathbb{R}$, $N \in (1, \infty]$. Let $U \in \mathcal{DC}_N$, let*

μ and σ be two elements of $P_2^{\rm ac}(M)$, let ρ be the density of μ with respect to ν, let ψ be a $d^2/2$-convex function such that $T = \exp(\widetilde{\nabla}\psi)$ is the unique Monge transport $\mu \to \sigma$, and for $s \in [0,1]$ let $\mu_s = (\exp(s\widetilde{\nabla}\psi))_\# \mu$. If $K < 0$ and $N < \infty$, further assume that the density ρ_s of μ_s is uniformly bounded below by a positive constant for small s:

$$\exists s_0 > 0; \quad \inf_{x \in M} \inf_{0 \le s \le s_0} \rho_s(x) > 0;$$

and that $p(r) = O(r^{1-\frac{1}{N}})$ as $r \to \infty$. Then

$$\lim_{s \downarrow 0} \frac{U_\nu(\mu_s) - U_\nu(\mu)}{s} = -\int p(\rho)\, L\psi \, d\nu,$$

where the function $L\psi$ is obtained from the measure $L\psi$ (understood in the sense of distributions) by keeping only the absolutely continuous part with respect to the volume measure.

Remark 23.12. The lower bound assumption on ρ_s in the case $K < 0$, $N < \infty$ is certainly too strong, but it is not clear to me how to relax it.

Proof of Theorem 23.11. **Step 1:** I shall prove the theorem in the case when the optimal map is of the form $\exp(\nabla\psi)$ (this is a true gradient, not an approximate gradient).

First note that ρ_0 is the same as ρ. The proof starts as in the proof of Theorem 17.15, with a change of variables:

$$\begin{aligned} U_\nu(\mu_s) = \int U(\rho_s(x))\, d\nu(x) &= \int U\Big(\rho_s(\exp_x s\nabla\psi(x))\Big)\, \mathcal{J}_{0\to s}(x)\, d\nu(x) \\ &= \int U\left(\frac{\rho_0(x)}{\mathcal{J}_{0\to s}(x)}\right) \mathcal{J}_{0\to s}(x)\, d\nu(x), \end{aligned}$$

where $\mathcal{J}_{0\to s}$ is the Jacobian determinant associated with the map $\exp(s\nabla\psi)$, and the reference measure ν. **Note that** here I am using the Jacobian formula for a change of variables which a priori is not Lipschitz. Upon use of $\mu = \rho_0\, \nu$, it follows that

$$\frac{U_\nu(\mu_s) - U_\nu(\mu)}{s} = \int \frac{1}{s}\left[U\left(\frac{\rho_0(x)}{\mathcal{J}_{0\to s}(x)}\right)\frac{\mathcal{J}_{0\to s}(x)}{\rho_0(x)} - \frac{U(\rho_0(x))}{\rho_0(x)}\right] d\mu(x). \tag{23.10}$$

By Theorem 14.1, for almost all x we have the Taylor expansion

$$\mathcal{J}_{0\to s}(x) = 1 + s\,(L\psi)(x) + o(s) \qquad \text{as } s \to 0. \tag{23.11}$$

On the other hand, for given r, the derivative of $\delta \to (\delta/r)\, U(r/\delta)$ at $\delta = 1$ is $-\,p(r)/r$. This and (23.11) imply that for almost all x where $\rho_0(x) > 0$,

$$\lim_{s\downarrow 0}\left[U\left(\frac{\rho_0(x)}{\mathcal{J}_{0\to s}(x)}\right)\frac{\mathcal{J}_{0\to s}(x)}{\rho_0(x)} - \frac{U(\rho_0(x))}{\rho_0(x)}\right] \;=\; -\,(L\psi(x))\left(\frac{p(\rho_0(x))}{\rho_0(x)}\right).$$

Then the conclusion follows after integrating against $d\mu(x) = \rho_0(x)\, d\nu(x)$.

So it only suffices to check that one can indeed pass to the limit as $s \to 0$ in (23.10). To do so, let

$$w(s,x) = \left[U\left(\frac{\rho_0(x)}{\mathcal{J}_{0\to s}(x)}\right)\frac{\mathcal{J}_{0\to s}(x)}{\rho_0(x)} - \frac{U(\rho_0(x))}{\rho_0(x)}\right];$$

the problem is to justify

$$\lim_{s\downarrow 0} \int w(s,x)\, d\mu(x) = \int \Big[\lim_{s\downarrow 0} w(s,x)\Big]\, d\mu(x). \tag{23.12}$$

Recall that

$$w(s,x) = \frac{u(s,x) - u(0,x)}{s}, \qquad u(s,x) = U\left(\frac{\rho_0(x)}{\mathcal{J}_{0\to s}(x)}\right)\frac{\mathcal{J}_{0\to s}(x)}{\rho_0(x)}.$$

First consider the case $K = 0$, which is simpler. From the estimates in the proof of Theorem 17.15 (recall (17.13)) we know that $u(s,x)$ is a convex function of s, for given x; then $w(s,x)$ is nonincreasing as $s \downarrow 0$, and (23.12) follows from the monotone convergence theorem.

Now consider the general case when $K < 0$. As in the estimates in the proof of Theorem 17.15 (see (17.13) again),

$$\frac{d^2\, u(s,x)}{ds^2} u(s,x) \geq K\lambda\, \rho_s\big(T_{0\to s}(x)\big)^{-\frac{1}{N}} \big|\nabla\psi_s(T_{0\to s}(x))\big|^2,$$

where $\nabla\psi_s$ gives the optimal velocity field to transport μ_s to σ. From our assumptions, $\rho_s^{-\frac{1}{N}}$ is bounded above; on the other hand, $|\nabla\psi_s(T_{0\to s}(x))| = (1-s)\, d(x, T(x))$, so there is a constant C such that

$$-\frac{d^2}{ds^2} u(s,x) \leq C\, d(x,T(x))^2.$$

Let then

$$\widetilde{u}(s,x) = u(s,x) + C\, d(x,T(x))^2\, s^2;$$

this is a convex function of s, and the reasoning which we used for $K = 0$ applies for u replaced by $\widetilde{u}$, so

$$\lim_{s\downarrow 0} \int \widetilde{w}(s,x)\, d\mu(x) = \int \Big[\lim_{s\downarrow 0} \widetilde{w}(s,x)\Big]\, d\mu(x), \tag{23.13}$$

where

$$\widetilde{w}(s,x) = \frac{\widetilde{u}(s,x) - \widetilde{u}(0,x)}{s} = w(s,x) - s\, C\, d(x,T(x))^2.$$

Since $\int d(x,T(x))^2\, d\mu(x) < +\infty$,

$$\lim_{s\to 0}\ s\, C \int d(x,T(x))^2\, d\mu(x) = 0,$$

so we can replace $\widetilde{w}$ by w in both sides of (23.13). This concludes the argument.

Step 2: Now let us consider the general case where the optimal map takes the form $\exp(\widetilde{\nabla}\psi)$, not $\exp(\nabla\psi)$. (The reader can skip this bit at first reading and go directly to the next theorem.) **asdf – REPRENDRE ICI** □

Theorem 23.11 has the merit to be exact: it provides the precise value of the derivative at $s = 0$. Unfortunately, the result is not in a form that can be readily exploited in conjunction with Theorem 23.8. The following technical Proposition will remedy this.

Theorem 23.13 (integration by parts). *Let M be a smooth Riemannian manifold, equipped with a reference measure $\nu = e^{-V}\mathrm{vol}$, $V \in C^2(M)$. Let $\psi : M \to \mathbb{R} \cup \{+\infty\}$ be a semi-convex function with a quadratic modulus, and let ζ be a C^1 nonnegative function on M supported in the interior of the domain of ψ. Then*

$$\int_M (L\psi)\, \zeta\, d\nu \leq -\int_M \langle \nabla\psi, \nabla\zeta\rangle\, d\nu. \tag{23.14}$$

In particular, if $U \in C^2(0,+\infty) \cap C(\mathbb{R}_+)$ is a given nonlinearity satisfying $U(0) = 0$, $p(r) = rU'(r) - U(r)$ is the associated pressure, ρ is a C^1 probability density on M, and either p is continuously differentiable at 0, or ρ is uniformly positive, then

$$\int (L\psi)\, p(\rho)\, d\nu \le -\int \big\langle \nabla\psi, \nabla p(\rho)\big\rangle\, d\nu.$$

Proof. **À RÉÉCRIRE. Y A-T-IL EGALITE PARFOIS ????** □

Diffusion equations as gradient flows

Let M be a Riemannian manifold **asdf hypothèses sur la courbure sectionnelle ??** satisfying the $\mathrm{CD}(K,N)$ curvature-dimension bound, and let U be some nonlinearity in $\mathcal{DC}_N$. By Theorem 17.15, U_ν is λ-displacement convex for some $\lambda \in \mathbb{R}$. Let then $(\mu_t)_{t\ge 0}$ be an absolutely continuous curve valued in $P_2^{\mathrm{ac}}(M)$, satisfying a continuity equation of the form

$$\frac{\partial \mu}{\partial t} + \nabla\cdot(\xi_t \mu_t) = 0,$$

where ξ is a Lipschitz vector field. Fix some time t, and let $\sigma \in P_2^{\mathrm{ac}}(M)$. Let $\exp(\nabla\psi)$ be the optimal transport $\mu_t \to \sigma$. Let also $\mu^{(s)} = \exp(s\nabla\psi)_\# \mu_t$. Then, under adequate regularity assumptions,

(a) according to Theorem 23.8,

$$\frac{d^+}{dt}\left(\frac{W_2(\sigma,\mu_t)^2}{2}\right) \le -\int \langle \nabla\psi, \xi_t\rangle\, d\mu_t;$$

(b) according to Theorems 23.11 and 23.13,

$$\frac{d^+}{ds}\, U_\nu(\mu^{(s)}) \le \int \big\langle \nabla\psi, \nabla p(\rho_t)\big\rangle\, d\mu_t.$$

So Proposition 23.6(i) applies as soon as

$$\xi = -\nabla p(\rho),$$

and guarantees that (μ_t) is an integral curve of the gradient flow associated with U_ν. The next theorem summarizes this reasoning:

Theorem 23.14 (Diffusion equations as gradient flows in Wasserstein space). *Let M be a smooth Riemannian manifold* **asdf hypothèses sur la courbure sectionnelle ?** *equipped with a reference measure $\nu = e^{-V}\mathrm{vol}$, $V \in C^2(M)$, satisfying the $\mathrm{CD}(K,N)$ curvature-dimension bound for some $K \in \mathbb{R}$, $N \in (1,\infty]$. Let $L = \Delta - \nabla V\cdot\nabla$. Let U be a nonlinearity in $\mathcal{DC}_N$, such that $U \in C^3(0,+\infty)$; and let $p(r) = rU'(r) - U(r)$. Let $\rho_t(x)$ be a (C^1 in t, C^2 in x) positive solution of the partial differential equation*

$$\frac{\partial \rho_t}{\partial t} = L p(\rho_t), \tag{23.15}$$

and let $\mu_t = \rho_t\, \nu$. If $K < 0$ and $N < \infty$, further assume that ρ_t is uniformly bounded below. Then $(\mu_t)_{t\ge 0}$ defines a trajectory ($\mu_t = \rho_t\,\nu$) of the gradient flow associated with the energy functional U_ν in $P_2^{\mathrm{ac}}(M)$.

Example 23.15. Any smooth positive solution of $\partial_t \rho = \Delta \rho$ can be seen as a trajectory of the gradient flow associated with the energy $H(\mu) = \int \rho \log \rho$. Any smooth positive solution of $\partial_t \rho = \Delta \rho + \nabla \cdot (\rho \nabla V)$ can be seen as a trajectory of the gradient flow associated with the energy $F(\mu) = \int \rho \log \rho + \int \rho V$. (With respect to the previous example, this amounts to change the reference measure vol into e^{-V} vol.) In nonnegative curvature, any smooth positive solution of $\partial_t \rho = \Delta \rho^m$, $m \geq 1 - 1/n$, can be seen as a trajectory of the gradient flow associated with the energy $E(\mu) = (m-1)^{-1} \int \rho^m$.

Proof of Theorem 23.14. Since M is compact and ρ is smooth positive, there is a lower bound r_0 for $\rho_t(x)$. Because U is C^3 on $(0, +\infty)$, the function $U'(\rho)$ is C^2, so

$$\xi_t(x) := -\nabla U'(\rho_t(x))$$

is a C^1 vector field. Then (23.15) can be rewritten as

$$\frac{\partial \rho_t}{\partial t} + \nabla \cdot (\xi_t \, \rho_t) = 0.$$

The theorem follows by the combination of Theorems 23.8, 23.11 and 23.13, as explained at the beginning of this section. □

In Theorem 23.14 I assumed the smoothness of the density; but in many situations there are regularization theorems for such (a priori nonlinear) diffusion equations, so the smoothness assumption can be relaxed in the end. Such is the case for the heat equation. Here is a result about this case, stated without proof:

Corollary 23.16 (Heat equation as a gradient flow). *Let M be a smooth Riemannian manifold* **asdf hypothèses sur la courbure sectionnelle ?** *curvature, let $V \in C^2(M)$, and let $L = \Delta - \nabla V \cdot \nabla$. Let $\mu_0 \in P_2(M)$, and let $\mu_t = \rho_t \text{vol}$ solve*

$$\frac{\partial \rho_t}{\partial t} = L \, p(\rho_t).$$

Then $(\mu_t)_{t>0}$ is a trajectory of the gradient flow associated with the energy functional

$$H_\nu(\mu) = \int \rho \log \rho \, d\nu, \qquad \mu = \rho \, \nu$$

in the Wasserstein space $P_2^{\text{ac}}(M)$.

In particular, the gradient flow associated with H_{vol} is the standard heat equation

$$\frac{\partial \rho}{\partial t} = \Delta \rho.$$

Remark 23.17. The distinction between $P_2(M)$ and $P_2^{\text{ac}}(M)$ is not essential here.

Remark 23.18. asdf hypothèses sur la courbure toujours en jeu ? éventuellement supprimer cette remarque The natural curvature assumption to define the heat equation is a lower bound on the *Ricci* curvature. So the question naturally arises whether the assumption of bounded sectional curvature can be relaxed. Also one might think about the role that could be played by global regularity results on the logarithm of the solution of the heat equation; see for instance Remark 24.17 below.

As I already said in the beginning of this chapter, the heat equation can be seen as a gradient flow in various ways. For instance, take for simplicity the basic heat equation in $\mathbb{R}^n$, in the form $\partial_t u = \Delta u$, then it can be interpreted as the gradient flow of the functional $E(u) = (1/2) \int |\nabla u|^2$ for the usual Hilbert structure imposed by the L^2 norm; or as as the gradient flow of the functional $E(u) = \int u^2$ for the Hilbert structure induced by the H^{-1} norm (say on the subspace $\int u = 0$). But the interesting new feature coming from optimal transport theory is that now the heat equation can be seen as the gradient flow of a nice functional which has statistical (or thermodynamical) meaning; and in such a way that it is naturally set in the space of probability measures.

General theory and time-discretization

There is a general theory of gradient flows in metric spaces, based for instance on Definition 23.5, or other variants appearing in Proposition 23.1. It was pushed to a high degree of sophistication by De Giorgi and his school, and other researchers. A key role in this theory is played by discrete-time **approximation schemes**, the simplest of which can be stated as follows:

1. Choose your initial datum X_0;

2. Choose a time step τ, which in the end will decrease to 0;

3. Construct $X_1^{(\tau)}$ as a mininimizer of $X \longmapsto \Phi(X) + \dfrac{d(X_0, X)^2}{2\tau}$; then construct inductively $X_{k+1}^{(\tau)}$ as a minimizer of $X \longmapsto \Phi(X) + \dfrac{d(X_k^{(\tau)}, X)^2}{2\tau}$.

4. Pass to the limit in $X_k^{(\tau)}$ as $\tau \to 0$, $k\tau \to t$, hopefully recover a function $X(t)$ which is the value of the gradient flow at time t.

Such schemes sometimes provide an excellent way to construct the gradient flow, and they may also be useful in numerical simulations. There are strong results about the convergence of such schemes; see the bibliographical notes for details.

This procedure also suggests a better intuition for the gradient flow in Wasserstein distance, as I shall explain in a slightly informal way. Consider, as in Theorem 23.14, the equation

$$\frac{\partial \rho}{\partial t} = Lp(\rho).$$

Suppose you know the density $\rho(t)$ at some time t, and look for the density $\rho(t + dt)$ at a later time, where dt is infinitesimally small. To do this, minimize the quantity

$$U_\nu(\mu_{t+dt}) - U_\nu(\mu_t) + \frac{W_2(\mu_t, \mu_{t+dt})^2}{2\, dt}.$$

There is another way to rewrite this, by using the interpretation of the Wasserstein distance between two infinitesimally close probability measures:

$$\frac{W_2(\mu_t, \mu_{t+dt})^2}{dt} \simeq \inf \left\{ \int |v|^2 \, d\mu_t; \quad \frac{\partial \mu}{\partial t} + \nabla \cdot (\mu v) = 0 \right\}.$$

All in all, to go from $\mu(t)$ to $\mu(t+dt)$, what you have to do is *find a velocity field v inducing an infinitesimal variation $d\mu = -\nabla \cdot (\mu v)\, dt$, so as to minimize the infinitesimal quantity*

$$dU_\nu + K\, dt, \tag{23.16}$$

where $U_\nu(\mu) = \int U(\rho)\, d\nu$, and K is the kinetic energy $(1/2)\int |v|^2\, d\mu$ (so $K\, dt$ is the infinitesimal action).

For the heat equation $\frac{\partial \rho}{\partial t} = \Delta \rho$, $\nu = \mathrm{vol}$, $U_\nu(\mu) = \int \rho \log \rho\, d\nu$, we are back to the example discussed in the beginning of this chapter.

There is an important moral here: Behind many *nonequilibrium* equations of statistical mechanics, there is a variational principle involving entropy and energy, or functionals alike, just as in equilibrium statistical mechanics.

Appendix: A lemma about doubling variables

The following important lemma was used in the proof of Theorems 22.34 and 23.8.

Lemma 23.19 (Differentiation through doubling of variables). *Let $F : [0,T] \times [0,T] \to \mathbb{R}$ be locally absolutely continuous in s, uniformly in t; and locally absolutely continuous in t, uniformly in s. Then $t \to F(t,t)$ is absolutely continuous, and for almost all t_0,*

$$\left.\frac{d}{dt}\right|_{t=t_0} F(t,t) = \limsup_{t \uparrow t_0} \left(\frac{F(t,t_0) - F(t_0,t_0)}{t - t_0} \right) + \limsup_{t \downarrow t_0} \left(\frac{F(t_0,t) - F(t_0,t_0)}{t - t_0} \right). \quad (23.17)$$

Explicitly, to say that F is locally absolutely continuous in s, uniformly in t, means that there is a fixed function $u \in L^1_{\mathrm{loc}}(dt)$ such that

$$\sup_{0 \le t \le T} \big| F(s,t) - F(s',t) \big| \le \int_s^{s'} u(\tau)\, d\tau.$$

Proof of Lemma 23.19. By assumption there are functions $u \in L^1_{\mathrm{loc}}(dt)$ and $v \in L^1_{\mathrm{loc}}(ds)$ such that

$$\begin{cases} \sup_{0 \le t \le T} \big| F(s,t) - F(s',t) \big| \le \int_s^{s'} u(\tau)\, d\tau \\ \sup_{0 \le t \le T} \big| F(s,t) - F(s',t) \big| \le \int_s^{s'} v(\tau)\, d\tau. \end{cases}$$

Without loss of generality we may take $u = v$.

Let $f(t) = F(t,t)$. Then

$$|f(s) - f(t)| \le |F(s,s) - F(s,t)| + |F(s,t) - F(t,t)| \le 2 \int_s^t u(\tau)\, d\tau;$$

so f is locally absolutely continuous.

Let $\dot f$ stand for the derivative of f. Since f is absolutely continuous, this is also (almost everywhere) the distributional derivative of f. The goal is to show that $\dot f(t)$ is bounded above by the right-hand side of (23.17).

Let ζ be a C^∞ nonnegative function supported in $(0,1)$. For h small enough, $\zeta(\cdot + h)$ is also supported in $(0,1)$, and

$$\int \dot f \zeta = -\int f \dot\zeta = \lim_{h \to 0} \int_0^1 f(t) \left[\frac{\zeta(t-h) - \zeta(t)}{h} \right] dt$$
$$= \lim_{h \to 0} \int_0^1 \zeta(t) \left[\frac{f(t+h) - \zeta(t)}{h} \right] dt.$$

Replacing f by its expression in terms of F, we get

$$\int \dot{f}\zeta = \lim_{h\downarrow 0}\left\{\int_0^1 \zeta(t)\left[\frac{F(t+h,t+h)-F(t,t+h)}{h}\right]dt + \int_0^1 \zeta(t)\left[\frac{F(t,t+h)-F(t,t)}{h}\right]dt\right\} \tag{23.18}$$

$$\leq \limsup_{h\downarrow 0}\int_0^1 \zeta(t-h)\left[\frac{F(t,t)-F(t-h,t)}{h}\right]dt + \limsup_{h\downarrow 0}\int_0^1 \zeta(t)\left[\frac{F(t,} \tag{23.19}$$

In the first integral on the right-hand side of (23.19), it is possible to replace $\zeta(t-h)$ by $\zeta(t)$, since

$$\begin{aligned}&\left|\int_0^1 \big[\zeta(t-h)-\zeta(t)\big]\left(\frac{F(t,t)-F(t-h,t)}{h}\right)dt\right| \\ &\leq (2\sup\zeta)\int_0^1 |F(t,t)-F(t-h,t)|\,dt \leq (2\sup\zeta)\int_0^1\left(\int_{t-h}^t u(\tau)\,d\tau\right)dt \\ &= (2\sup\zeta)h\int_0^1 u(\tau)\,d\tau = O(h).\end{aligned}$$

To summarize:

$$\int_0^1 \zeta\dot{f} \leq \limsup_{h\downarrow 0}\int_0^1 \zeta(t)\left[\frac{F(t,t)-F(t-h,t)}{h}\right]dt + \limsup_{h\downarrow 0}\int_0^1 \zeta(t)\left[\frac{F(t,t_h)-F(t,t)}{h}\right]dt. \tag{23.20}$$

By assumption,

$$\left|\frac{F(t,t)-F(t-h,t)}{h}\right| \leq \frac{1}{h}\int_{t-h}^t u(\tau)\,d\tau;$$

and by Lebesgue's density theorem, the right-hand side converges in $L^1_{\mathrm{loc}}(dt)$ as $h\to 0$. This makes it possible to apply Fatou's lemma, in the form

$$\limsup_{h\downarrow 0}\int_0^1 \zeta(t)\left[\frac{F(t,t)-F(t-h,t)}{h}\right]dt \leq \int_0^1 \zeta(t)\,\limsup_{h\downarrow 0}\left[\frac{F(t,t)-F(t-h,t)}{h}\right]dt. \tag{23.21}$$

Similarly,

$$\limsup_{h\downarrow 0}\int_0^1 \zeta(t)\left[\frac{F(t,t+h)-F(t,t)}{h}\right]dt \leq \int_0^1 \zeta(t)\,\limsup_{h\downarrow 0}\left[\frac{F(t,t+h)-F(t,t)}{h}\right]dt. \tag{23.22}$$

Plugging (23.21) and (23.22) back in (23.20), we find that

$$\int_0^1 \zeta\dot{f} \leq \int \zeta(t)\left\{\limsup_{h\downarrow 0}\left(\frac{F(t,t)-F(t-h,t)}{h}\right)+\limsup_{h\downarrow 0}\left(\frac{F(t,t+h)-F(t,t)}{h}\right)\right\}dt.$$

Since ζ is arbitrary, $\dot{f}$ is bounded above by the expression in curly brackets, almost everywhere. This concludes the proof. □

Bibliographical Notes

Historically, the development of the theory of abstract gradient flows was impulsed by De Giorgi on the basis of the time-discretized variational scheme, and by Bénilan [45] on the basis of the variational inequalities involving the square distance, as in Proposition 23.1(iv). The latter approach has the advantage to incorporate stability and uniqueness as a built-in feature, while the former is more efficient in establishing existence. Bénilan introduced his method in the setting of Banach spaces, but it works just as well in abstract metric spaces. Luckily, both approaches work in the Wasserstein space.

Currently, **the** reference for abstract gradient flows is the recent monograph by Ambrosio, Gigli and Savaré [15]; there is a short version in Ambrosio's Santander lecture notes [11]. There the reader will find the most precise results known to this day, apart from some very recent refinements which are due to the same authors. (For instance, Ambrosio and Savaré recently found out a simplified proof of error estimates and convergence proofs for time-discretized gradient flows.) More than half of the book is devoted to gradient flows in the space of probability measures on $\mathbb{R}^n$ (or a separable Hilbert space). Issues about the replacement of $P_2(\mathbb{R}^n)$ by $P_2^{\mathrm{ac}}(\mathbb{R}^n)$ are also carefully discussed there. It would be great if someone had the courage to adapt all the results in this book to general smooth Riemannian manifolds rather than just $\mathbb{R}^n$.

The classical theory of gradient flows in Hilbert spaces, mostly for convex functionals, is developed in Brézis [78] and other sources; it is also implicitly used in several parts of the popular book by J.-L. Lions [241].

The differentiability of the Wasserstein distance in $P_2^{\mathrm{ac}}(\mathbb{R}^n)$, and in fact in $P_p^{\mathrm{ac}}(\mathbb{R}^n)$ ($1 < p < \infty$), is proven in [15, Theorems 10.2.2 and 10.2.6, Corollary 10.2.7]. The assumption of absolute continuity of the probability measures is not crucial for the super-differentiability (actually in [15, Theorem 10.2.2] there is no such assumption). For the sub-differentiability, this assumption is only used to guarantee the uniqueness of the transference plan.

For many applications however, there is a more relevant and general statement that the Wasserstein distance $W_2(\sigma, \mu_t)$ is almost surely (in t) differentiable along any absolutely continuous curve $(\mu_t)_{0 \le t \le 1}$; this holds true without any assumption of absolute continuity of the measures [15, Theorem 8.4.7]. (But in this reference, only the Euclidean space is considered.)

About Theorem 23.11 and Proposition 23.13, the remarks made in the first paragraph of the bibliographical notes for Chapter 20 also apply here. Theorem 23.11 is, to my knowledge, the first result of its kind in a truly Riemannian setting. There are several technical assumptions (lower bound on ρ in case of negative curvature, **asdf bornes sur la courbure sectionnelle ?**) which one might wish to relax.

Theorem 23.13 was first established in $\mathbb{R}^n$, independently in [102, Lemma 5.12] and [15, Lemma 10.4.5] by very slightly different arguments. A particular case (with power laws nonlinearities) was first considered in [119], in relation with the discussion of equality cases in optimal Sobolev inequalities.

The interpretation of the linear Fokker–Planck equation $\partial_t \rho = \Delta \rho + \nabla \cdot (\rho \nabla V)$ as the limit of a discretized scheme goes back to the pioneering work of Jordan, Kinderlehrer and Otto [219]. In that sense the Fokker–Planck equation can be considered as the abstract gradient flow corresponding to the free energy $\Phi(\rho) = \int \rho \log \rho + \int \rho V$. The proof (slightly rewritten) appears in my book [365, Section 8.5]. It is based on the three main estimates which are more or less at the basis of the whole theory of abstract gradient flows: if τ is the time step, and $X_k^{(\tau)}$ the position at step k of the discretized system, then

$$\begin{cases} d(X_k^{(\tau)}, X_0) = O(1), \\ \\ \displaystyle\sum_{j=1}^{\infty} \frac{d\big(X_j^{(\tau)}, X_{j+1}^{(\tau)}\big)^2}{2\tau} = O(1), \\ \\ \Phi(X_0) - \Phi(X_{n+1}^{(\tau)}) = O(1). \end{cases}$$

Here I have assumed that Φ is bounded below (which is the case when Φ is the free energy functional). When Φ is not bounded below, there are estimates of the same type, but quite more complicated [15, Section 3.2].

Otto applied the same method to various classes of nonlinear diffusion equations, including porous medium and fast diffusion equations [290], and parabolic p-Laplace type equations [287], but also more exotic models [288, 289]. For background about the theory of porous medium and fast diffusion equations, the reader may consult the review texts by Vázquez [363, 362].

In his work about porous medium equations, Otto introduced the abstract formalism allowing him to interpret these equations as gradient flows, directly at the continuous level (without going through the time-discretization). The psychological impact of this work on specialists of optimal transport was important. Otto's approach was developed by various authors, including Carrillo, McCann and myself [103, 102] and Ambrosio, Gigli and Savaré [15].

The setting adopted in [365, 102, 15] is the following: Let E denote an energy functional of the form

$$E(\mu) = \int_{\mathbb{R}^n} U(\rho(x))\, dx + \int_{\mathbb{R}^n} V(x)\, d\mu(x) + \frac{1}{2} \int_{\mathbb{R}^n \times \mathbb{R}^n} W(x-y)\, d\mu(x)\, d\mu(y),$$

where as usual ρ is the density of μ, and $U(0) = 0$; then under certain regularity assumptions, the associated gradient flow with respect to the 2-Wasserstein distance W_2 is

$$\frac{\partial \rho}{\partial t} = \Delta p(\rho) + \nabla \cdot (\rho\, \nabla V) + \nabla \cdot \big(\rho\, \nabla(\rho * W)\big),$$

where as usual $p(r) = rU'(r) - U(r)$. (When $p(r) = r$, the above equation is a special case of McKean–Vlasov equation.) The most general results of this kind can be found in [15]. Such equations arise in a number of physical models; see e.g. [102].

Other interesting gradient flows are obtained by choosing for the energy functional

- the Fisher information

$$I(\mu) = \int \frac{|\nabla \rho|^2}{\rho};$$

then the resulting equation is the quantum drift-diffusion equation [15, Example 11.1.10]. This was recently studied at a rigorous level by Gianazza, Savaré and Toscani [184].

- the squared H^{-1} norm

$$\|\mu\|_{H^{-1}}^2 = \|\nabla \Delta^{-1} \rho\|_{L^2}^2;$$

then the resulting equation appears in the Ginzburg-Landau dynamics. This idea has been in the air for a few years at a purely formal level; recently, Ambrosio and Serfaty [**?**] have made some preliminary progress on its rigorous justification.

Gradient flows with respect to the Wasserstein distances W_p with $p \neq 2$ were considered in [287] and lead to other classes of well-known diffusion equations, such as p-Laplace

equations $\partial_t \rho = \nabla \cdot (|\nabla \rho|^{p-2} \nabla \rho)$. A large part of the discussion can be transposed to that case [2, 256], but things become quite more difficult.

Brenier [77] has suggested that certain cost functions with "relativistic" features could be physically relevant, for instance $c(x, y) = c(x - y)$ with

$$c(v) = 1 - \sqrt{1 - \frac{|v|^2}{c^2}} \qquad \text{or} \qquad c(v) = \sqrt{1 + \frac{|v|^2}{c^2}} - 1.$$

By applying the general formalism of gradient flows with such cost functions, he derived relativistic-like heat equations, such as

$$\frac{\partial \rho}{\partial t} = \nabla \cdot \left(\frac{\rho \nabla \rho}{\sqrt{\rho^2 + \varepsilon^2 |\nabla \rho|^2}} \right).$$

This looked a bit like a formal game, but it was later found out that related equations were common in the physical literature about flux-limited diffusion processes [273], and that in fact Brenier's very equation had already been considered by Rosenau [313]. A rigorous treatment of these equations leads to challenging analytical difficulties, which triggered several recent technical works, see e.g. [20, 19] and the references therein.

A few comments are in order about the appearance of variational problems in statistical mechanics. For the most part, equilibrium statistical mechanics rests on the idea that the equilibrium measure is obtained by the minimization of a thermodynamical functional such as the free energy. The principle according to which *nonequilibrium* statistical mechanics may also be understood through variational principles is much more original; I first heard it explicitly in a talk by Kinderlehrer (June 1997 in Paris), about the interpretation of the Fokker–Planck equation by means of Wasserstein distance. Independently of optimal transport theory, the same idea has been making its way in the community of physicists, where it may be attributed to Prigogine. There is ongoing research in that direction, in relation to large deviations and fluctuation of currents, performed by Gabrieli, Landim, Derrida, Lebowitz, Speer, Jona Lasinio and others. It seems to me that both approaches (optimal transport on one hand, large deviations on the other) have a lot in common, although the formalisms look very different. By the way, some links between optimal transport and large deviations have recently been explored in a book by Feng and Kurtz [163].

So far I have mainly discussed gradient flows associated with cost functions that are quadratic ($p = 2$), or at least strictly convex. But there are some quite interesting models of gradient flows for, say, the cost function which is equal to the distance ($p = 1$). Such equations have been used for instance in the modelling of sandpiles [304, 26, 155, 159, 28], or compression molding [27]. These issues are briefly reviewed by Evans [154].

Also very recently, variational problems taking the form of a discretized gradient flow have made their way in mathematical economics or decision theory; in these models the negative of the energy can be thought of as, say, the reward or the benefits obtained from a certain skill or method or decision, while the cost function can be interpreted as the effort or difficulty which one has to spend in order to learn this skill or change one's habits or take the decision. As an entry point to that literature, the reader may take a look at a paper by Attouch and Soubeyran [29]. It is interesting to note that the gradient flows in this kind of literature would rather be of the kind $p = 1$ than of the kind $p = 2$.

This chapter was only concerned with gradient flows. The situation concerning Hamiltonian flows is anything but clear. In [365, Section 8.3.2] one can find some examples of equations that one would like to intepret as Hamiltonian equations with respect to the distance W_2, and other equations that one would like to interpret as dissipative Hamiltonian

equations. Another example is the rescaled two-dimensional incompressible Navier-Stokes equation in vorticity formulation (for nonnegative vorticity), as studied by Gallay and Wayne [174]. As far as rigorous justification is concerned, there is a recent paper by Ambrosio and Gangbo [14] which covers certain classes of Hamiltonian equations, yet not as wide as one could wish.

A particularly interesting "dissipative Hamiltonian equation" that should have an interpretation in terms of optimal transport is the kinetic Fokker–Planck equation, with or without self-interaction. Huang and Jordan [217] studied this model, but it seems to me that their will to interpret it in the framework of gradient flows (rather than "dissipative Hamiltonian flows") led them to somewhat artificial rescalings. On this subject there is also a contribution by Carlen and Gangbo [99], with a completely different point of view. So far no clear picture has emerged, and the precise sense in which the word "Hamiltonian" should be interpreted has not yet been identified. According to a comment made to me by Ghys, it looks more like a Poisson structure rather a true Hamiltonian structure.

Lemma 23.19 in the Appendix is borrowed from [15, Lemma 4.3.4]. As Ambrosio pointed out to me, the argument is reminiscent of Kruzkhov's doubling method for the proof of uniqueness in the theory of scalar conservation laws, see for instance the nice presentation in [153, Sections 10.2 and 11.4].

24

Gradient flows II: Qualitative properties

Consider a smooth Riemannian manifold M, equipped with a reference measure $\nu = e^{-V}\mathrm{vol}$, and a partial differential equation such as

$$\frac{\partial \rho}{\partial t} = L\, p(\rho), \tag{24.1}$$

where $p(r) = rU'(r) - U(r)$, U is a given nonlinearity, the unknown $\rho = \rho(t,x)$ is a probability density on M and $L = \Delta - \nabla V \cdot$.

As I explained in Chapter 23, equation (24.1) can be interpreted as a gradient flow in the Wasserstein space $P_2(M)$. What do we gain from that information? Among other things, a set of recipes and estimates associated with gradient flows. This is what I shall illustrate in this chapter.

Here are the conventions to be used in this chapter:

• M is a Riemannian manifold, d is its geodesic distance and vol its volume;

• $\nu = e^{-V}\,\mathrm{vol}$ is a reference measure on M;

• $L = \Delta - \nabla V \cdot \nabla$ is a linear differential operator admitting ν for invariant measure;

• U is a convex nonlinearity with $U(0) = 0$; typically U will belong to some $\mathcal{DC}_N$ class;

• $p(r) = rU'(r) - U(r)$ is the pressure function associated to U;

• $\mu_t = \rho_t\, \nu$ is the solution of a certain partial differential equation $\partial_t \rho_t = L\, p(\rho_t)$ (sometimes I shall say that μ is the solution, sometimes that ρ is the solution);

• $U_\nu(\mu) = \int_M U(\rho)\, d\nu$; $\quad I_{U,\nu}(\mu) = \int_M \rho\, |\nabla U'(\rho)|^2\, d\nu$.

Calculation rules

Having put equation (24.1) in gradient flow form provides the possibility to use Otto's calculus to shortcut certain formal computations, and quickly get relevant results, without risks of errors of computations. When it comes to rigorous justification, things however are not so nice, and regularity issues (the daily nightmare of specialists of partial differential equations) should be considered. For the most important of these partial differential equations (heat or Fokker–Planck equation; porous medium equations) these regularity issues are nowadays under good control.

Examples 24.1. Consider a power law nonlinearity $U(r) = r^m$, $m > 0$. For $m > 1$ the resulting equation (24.1) is called a porous medium equation, and for $m < 1$ a fast diffusion equation. These equations are usually studied under the restriction $m > 1 - (2/n)$, because for $m \leq 1 - (2/n)$ the solution might fail to exist (there is in general loss of mass at infinity in finite time, or even in no time). If M is compact and ρ_0 is positive, then there is a unique

C^∞, positive solution. For $m > 1$, if ρ_0 vanishes somewhere, the solution in general fails to have C^∞ regularity at the boundary of the support of ρ. For $m < 1$, adequate decay conditions at infinity are needed.

To avoid considerably inflating the size of this paper, I shall **not** go into this problem in this chapter, and be content with theorems that will be conditional to the regularity of the solution.

Theorem 24.2 (computations for gradient flows). *Let $\rho(t,x)$ be a solution of (24.1) defined on $\mathbb{R}_+ \times M$. Let further A be a convex nonlinearity, C^2 on $[0,+\infty)$. Assume that*

(a) ρ is bounded and positive on $[0,T) \times M$, for any $T < \infty$;

(b) ρ is C^3 in the x variable and C^1 in the t variable;

(c) U is C^4 on $(0,+\infty)$;

(d) V is C^4 on M;

(e) For any $t > 0$,

$$\sup_{|s-t|<\delta} \frac{1}{|t-s|} \Big(|\rho_t(x) - \rho_s(x)| + |U(\rho_t(x)) - U(\rho_s(x))| + |LU'(\rho_t(x))\, p(\rho_t(x)) - LU'(\rho_s(x))\, p(\rho_s(x))| \Big) \in L^1(\nu(dx));$$

(f) ρ, $p(\rho)$, $Lp(\rho)$, $p_2(\rho)$, $\nabla p_2(\rho)$, $U'(\rho)$, $\nabla U'(\rho)$, $LU'(\rho)$, $\nabla LU'(\rho)$, $L|\nabla U'(\rho)|^2$, $L(\nabla U'(\rho)\, \nabla LU'(\rho))$ and e^{-V} satisfy adequate growth/decay conditions at infinity.

Then the following formulas hold true:

(i) $\forall t > 0, \quad \dfrac{d}{dt} \displaystyle\int A(\rho_t)\, d\nu = - \int p'(\rho_t)\, A''(\rho_t) |\nabla \rho_t|^2\, d\nu;$

(ii) $\forall t > 0, \quad \dfrac{d}{dt} U_\nu(\mu_t) = -I_{U,\nu}(\mu_t);$

(iii) $\forall t > 0, \quad \dfrac{d}{dt} I_{U,\nu}(\mu_t) = -2 \displaystyle\int_M \Big[\|\nabla^2 U'(\rho_t)\|^2_{\mathrm{HS}} + \big(\mathrm{Ric} + \nabla^2 V\big)(\nabla U'(\rho_t)) \Big] p(\rho) d\nu + \int_M \big(LU'(\rho_t)\big)^2 p_2(\rho_t)\, d\nu.$

Particular Case 24.3. In the particular case $U(r) = r \log r$, Formula (ii) is a famous identity: **the Fisher information is the time-derivative of the entropy along the heat semigroup.** (What I call entropy is not H_ν but $-H_\nu$; this coincides with the physicists' convention.)

In the sequel, what I call *smooth solution* of (24.1) is a solution satisfying Assumptions (a) to (f) above.

Remark 24.4. I have not been been precise about the conditions at infinity needed in Assumption (f), because there are a large number of possible assumptions. The point is to be able to justify a certain number of integrations by parts, using integrability and moment conditions. This is true for instance if $V = 0$, ρ, $p(\rho)$ and $p_2(\rho)$ have finite moments of all orders and $U'(\rho)$ and all its derivatives have polynomial growth. When V is not zero, there might be issues about the density of $C^\infty_c(M)$ in the weighted Sobolev spaces $H^1(e^{-V})$ and $H^2(e^{-V})$ which are associated with the operator L. These problems are worsened by the issues about the behavior of the manifold M at infinity.

Formal proof of Theorem 24.2. By Otto's calculus (Formula 15.2),

$$\begin{aligned}\frac{d}{dt}\int A(\rho_t)\,d\nu &= -\Big\langle \mathrm{grad}_{\mu_t} A_\nu,\ \mathrm{grad}_{\mu_t} U_\nu \Big\rangle \\ &= -\int \rho_t\,\nabla A'(\rho_t)\cdot\nabla U'(\rho_t)\,d\nu \\ &= -\int \rho_t U''(\rho_t)\,A''(\rho_t)\,|\nabla\rho_t|^2\,d\nu \\ &= -\int p'(\rho_t)\,A''(\rho_t)\,|\nabla\rho_t|^2\,d\nu.\end{aligned}$$

This leads to formula (i). The particular case when $A = U$ gives

$$\begin{aligned}\frac{d}{dt}\int U(\rho_t)\,d\nu &= -\Big\|\mathrm{grad}_{\mu_t} U_\nu\Big\|^2 \\ &= -\int \rho_t|\nabla U'(\rho_t)|^2\,d\nu = -I_{U,\nu}(\mu_t),\end{aligned}$$

which is (ii).

Finally, we can differentiate the previous expression once again along the gradient flow $\dot\mu = -\mathrm{grad}U_\nu$:

$$\frac{d}{dt}\Big\|\mathrm{grad}_{\mu_t} U_\nu\Big\|^2 = -2\Big\langle \mathrm{Hess}_{\mu_t}\cdot\mathrm{grad}_{\mu_t} U_\nu,\ \mathrm{grad}_{\mu_t} U_\nu\Big\rangle,$$

and then (iii) follows from Formula 15.7. □

Rigorous proof of Theorem 24.2. A crucial observation is that (24.1) can be rewritten $\partial_t\rho = \nabla\cdot(\rho_t\nabla U'(\rho_t))$, where $\nabla\cdot$ stands for the adjoint of the gradient operator *with respect to the reference measure* ν. (With respect to the usual divergence operator, there is an additional term $\nabla V\cdot\nabla$.) Then the proofs of (i) and (ii) are obtained by just repeating the arguments by which Formula 15.2 was established. This is a succession of differentiations under the integral, chain-rules and integrations by parts:

$$\begin{aligned}\frac{d}{dt}\int A(\rho_t)\,d\nu &= \int \partial_t[A(\rho_t)]\,d\nu \\ &= \int A'(\rho_t)\,(\partial_t\rho_t)\,d\nu \\ &= \int A'(\rho_t)\,\nabla(\rho_t\,\nabla U'(\rho_t))\,d\nu \\ &= -\int \nabla A'(\rho_t)\,\rho_t\,\nabla U'(\rho_t)\,d\nu,\end{aligned}$$

and then the rest of the computation is the same as before.

The justification of (iii) is more tricky. First write

$$\int \rho|\nabla U'(\rho)|^2\,d\nu = \int U'(\rho)\,\nabla\cdot(\rho\,\nabla U'(\rho))\,d\nu = \int U'(\rho)\,Lp(\rho)\,d\nu = \int LU'(\rho)\,p(\rho),$$

where the self-adjointness of L with respect to the measure ν was used. Then

$$\begin{aligned}\frac{d}{dt}\int LU'(\rho_t)\,p(\rho_t)\,d\nu &= \int \partial_t\big(LU'(\rho_t)\big)\,p(\rho_t)\,d\nu + \int LU'(\rho_t)\,p(\rho_t)\,\partial_t\big(p(\rho_t)\big)\,d\nu \\ &= \int L(\partial_t U'(\rho_t))\,p(\rho_t)\,d\nu + \int LU'(\rho_t)\,p'(\rho_t)\,\nabla\cdot(\rho_t\,\nabla U'(\rho_t))\,d\nu.\end{aligned} \tag{24.2}$$

On the other hand,

$$\begin{aligned}\partial_t U'(\rho_t) = U''(\rho_t)\,\partial_t\rho &= U''(\rho_t)\,\nabla\cdot(\rho_t\,\nabla U'(\rho_t))\\ &= U''(\rho_t)\,\nabla\rho_t\cdot\nabla U'(\rho_t) + \rho_t U''(\rho_t)\,LU'(\rho_t)\\ &= |\nabla U'(\rho_t)|^2 + \rho U''(\rho_t)\,LU'(\rho_t).\end{aligned}$$

Plugging this back in (24.2), we obtain

$$\begin{aligned}\frac{d}{dt}\int LU'(\rho_t)\,p(\rho_t)\,d\nu = \int L|\nabla U'(\rho_t)|^2\,p(\rho)\,d\nu + \int L\Big(\rho_t U''(\rho_t)\,LU'(\rho_t)\Big)\,d\nu\\ + \int LU'(\rho_t)p'(\rho_t)\,\nabla\cdot(\rho_t\,\nabla U'(\rho_t))\,d\nu. \quad (24.3)\end{aligned}$$

The last two terms in this formula are actually equal: Indeed,

$$\begin{aligned}\int L\Big(\rho U''(\rho)\,LU'(\rho)\Big)\,d\nu &= \int \rho U''(\rho)\,LU'(\rho)\,Lp(\rho)\,d\nu\\ &= \int p'(\rho)\,LU'(\rho)\,\nabla\cdot(\rho\,\nabla U'(\rho))\,d\nu.\end{aligned}$$

So the expression appearing in (24.3) is exactly *twice* the expression appearing in (15.18), up to the replacement of ψ by $-U'(\rho_t)$. To arrive at formula (iii), it suffices to repeat the computations leading from (15.18) to (15.20), and to apply Bochner's formula. □

Large-time behavior

Otto's calculus, described in Chapter 15, was first derived to estimate rates of equilibration for certain nonlinear diffusion equations. The next theorem illustrates this.

Theorem 24.5 (equilibration under a $\mathrm{CD}(K,N)$ condition). *Let M be a Riemannian manifold equipped with a reference measure $\nu = e^{-V}$, satisfying the curvature-dimension bound $\mathrm{CD}(K,N)$ for some $K > 0$, $N \in (1,\infty]$, and let $U \in \mathcal{DC}_N$. Then,*

(i) [exponential convergence to equilibrium] Any smooth solution $(\mu_t)_{t\geq 0}$ of (24.1) satisfies the following estimates:

$$\begin{cases}(a) \qquad [U_\nu(\mu_t) - U_\nu(\nu)] \leq e^{-2K\lambda t}\,[U_\nu(\mu_0) - U_\nu(\nu)];\\ (b) \qquad I_{U,\nu}(\mu_t) \leq e^{-2K\lambda t}\,I_{U,\nu}(\mu_0);\\ (c) W_2(\mu_t,\nu) \leq e^{-2K\lambda t}\,W_2(\mu_0,\nu),\end{cases} \quad (24.4)$$

where

$$\lambda := \left(\lim_{r\to 0}\frac{p(r)}{r^{1-\frac{1}{N}}}\right)\sup_{x\in M}\rho_0(x)^{\frac{1}{N}}. \quad (24.5)$$

In particular, λ is independent of ρ_0 if $N = \infty$.

(ii) [exponential contraction] Any two solutions $(\mu_t)_{t\geq 0}$ and $(\widetilde{\mu}_t)_{t\geq 0}$ of (24.1) satisfy

$$W_2(\mu_t,\widetilde{\mu}_t) \leq e^{-K\lambda t}W_2(\mu_0,\widetilde{\mu}_0), \quad (24.6)$$

where

$$\lambda := \left(\lim_{r\to 0}\frac{p(r)}{r^{1-\frac{1}{N}}}\right)\sup_{x\in M}\max(\rho_0(x)^{\frac{1}{N}},\,\rho_1(x)^{\frac{1}{N}}). \quad (24.7)$$

Example 24.6. Smooth solutions of the Fokker–Planck equation

$$\frac{\partial \rho}{\partial t} = L\rho \tag{24.8}$$

converge to equilibrium at speed at least as fast as $O(e^{-Kt})$, in W_2 distance, in entropy sense (i.e. in the sense of the convergence of $\sqrt{H_\nu(\mu)}$ to 0), and in Fisher information sense.

Remark 24.7. At least formally, these properties are in fact *general properties of gradient flows*: If F satisfy $\mathrm{Hess} F \geq \lambda > 0$, on a geodesically convex subset of a manifold, X_∞ is the minimizer of F and $X(t)$, $\widetilde{X}(t)$ are two trajectories of the gradient flow associated with F, then we have the three estimates $[F(X(t)) - F(X_\infty)] \leq e^{-\lambda t}[F(X(0)) - F(X_\infty)]$; $|\nabla F(X(t))| \leq e^{-\lambda t}|\nabla F(X(0))|$; $d(X(t), \widetilde{X}(t)) \leq e^{-\lambda t} d(X(0), \widetilde{X}(0))$. As a good exercise, the reader can try to prove these estimates directly.

Remark 24.8. The rate of decay $O(e^{-\lambda\, t})$ is optimal for (24.8) if dimension is not taken into account; but if N is finite, the optimal rate of decay is $O(e^{-\lambda t})$ with $\lambda = KN/(N-1)$. The method presented in this chapter is not clever enough to catch this sharp rate.

Remark 24.9. I believe that the preceding results of convergence are satisfactory as I have stated them, i.e. in terms of convergence of natural, physically meaningful functionals However, it is also often possible to get similar rates of decay for more classical distances such as the L^1 norm, thanks to the **Csiszár–Kullback–Pinsker inequality**

$$\big\|\mu - \nu\big\|_{TV} \leq \sqrt{2\, H_\nu(\mu)}, \tag{24.9}$$

and generalizations thereof.

Remark 24.10. If $N < \infty$, Theorem 24.5 proves convergence to equilibrium with a rate that depends on the initial datum. However, if the solution $(\rho_t)_{t\geq 0}$ satisfies *uniform* smoothness bounds and M is compact, then it is usually possible to reinforce the statement $\rho_t \underset{L^1}{\longrightarrow} 1$ into $\rho_t \underset{L^\infty}{\longrightarrow} 1$. Then we can choose ρ_T as new initial datum, and get

$$t \geq T \Longrightarrow \qquad U_\nu(\mu_t) \leq e^{-K\lambda_T\,(t-T)} U_\nu(\mu_T) \ \leq e^{-K\lambda_T\,(t-T)} U_\nu(\mu_0), \tag{24.10}$$

where $\lambda_T = (\lim p(r)/r^{1-1/N})\, \sup \rho_T(x)^{\frac{1}{N}}$ converges to $\lambda_\infty = (\lim p(r)/r^{1-1/N})$ as $T \to \infty$. It follows from (24.10) that μ_t converges to ν as $O(e^{-K\widetilde{\lambda}\, t})$ for any $\widetilde{\lambda} > \lambda$.

Proof of Theorem 24.5. Let $H(t) = U_\nu(\mu_t)$; by Theorem 24.2(ii), we have $H'(t) = -I_{U,\nu}(\mu_t)$. On the other hand, the (modified) Sobolev inequality of Theorem 21.7 reads

$$U_\nu(\mu_t) \leq \frac{(\sup \rho)^{\frac{1}{N}}}{2K\lambda}\, I_{U,\nu}(\mu_t).$$

Thus,

$$\frac{d}{dt} H(t) \leq -2K\lambda (\sup \rho_t)^{-1/N} H(t). \tag{24.11}$$

Theorem 24.2(i) with $A(r) = r^p$, $p \geq 2$, gives

$$\int \rho \nabla U'(\rho) \cdot \nabla L'(\rho)\, d\nu,$$

$$\frac{d}{dt}\int \rho^p \, d\nu = -p(p-1)\int \rho U''(\rho)\rho^{p-2}|\nabla\rho|^2 \, d\nu \le 0.$$

So $\|\rho_t\|_{L^p}^p$ is a nonincreasing function of t, and therefore

$$\forall t \ge 0 \qquad \|\rho_t\|_{L^p(\nu)} \le \|\rho_0\|_{L^p(\nu)}.$$

Passing to the limit as $p \to \infty$ yields

$$\forall t \ge 0 \qquad \sup \rho_t \le \sup \rho_0.$$

Plugging this back in (24.11), we get

$$\frac{d}{dt}H(t) \le -2K\lambda(\sup\rho_0)^{-1/N}H(t),$$

and then (24.4)(a) follows.

Next, if $U \in \mathcal{DC}_N$, and $\mathrm{CD}(K,N)$ is enforced, we can write, as in (16.11),

$$\begin{aligned}
-(1/2)\frac{d}{dt}\, I_{U,\nu}(\mu_t) &= \int_M \Gamma_2(U'(\rho_t))\, p(\rho_t)\, d\nu + \int_M (LU'(\rho_t))^2\, p_2(\rho_t)\, d\nu \\
&\ge \int_M \mathrm{Ric}_{N,\nu}(\nabla U'(\rho_t))\, p(\rho_t)\, d\nu + \int_M (LU'(\rho_t))^2 \left[p_2 + \frac{p}{N}\right](\rho_t)\, d\nu \\
&\ge K\int_M |\nabla U'(\rho_t)|^2\, p(\rho_t)\, d\nu + \int_M (LU'(\rho_t))^2 \left[p_2 + \frac{p}{N}\right](\rho_t)\, d\nu \\
&\ge K\int_M |\nabla U'(\rho_t)|^2\, p(\rho_t)\, d\nu \\
&\ge K\lambda(\max\rho_t)^{-\frac{1}{N}} \int_M |\nabla U'(\rho_t)|^2\, \rho_t\, d\nu \\
&= K\lambda(\max\rho_t)^{-\frac{1}{N}}\, I_{U,\nu}(\mu_t).
\end{aligned}$$

Then (24.4)(b) follows.

It remains to establish (24.6) (of which (24.4)(c) is obviously a corollary). Let $t > 0$ be given, and let **J'AI BESOIN D'HYPOTHÈSES SUR LA COURBURE POUR AVOIR LE VRAI GRADIENT ??.....** $T = \exp(\nabla\psi)$ (resp. $\widetilde{T} = \exp(\nabla\widetilde{\psi})$ be the unique optimal transport between μ_t and $\widetilde{\mu}_t$, with ψ (resp. $\widetilde{\psi}$) $d^2/2$-convex. Theorem 23.8 reads

$$\frac{d}{dt}W_2(\mu_t,\widetilde{\mu}_t)^2 \le 2\int \langle \nabla\psi\, \nabla U'(\rho_t)\rangle\, d\mu_t + 2\int \langle \nabla\widetilde{\psi},\, \nabla U'(\widetilde{\rho}_t)\rangle\, d\widetilde{\mu}_t. \tag{24.12}$$

Let then $(\mu_t^{(s)})_{0\le s\le 1}$ be the unique displacement interpolation connecting $\mu_t^{(0)} = \mu_t$ to $\mu_t^{(1)} = \widetilde{\mu}_t$. By Theorems 23.11 and 23.13,

$$\left.\frac{d}{ds}\right|_{s=0^+} U_\nu(\mu_t^{(s)}) = -\int p(\rho_t)\, Ł\psi\, d\nu = \int\int \langle \nabla\psi,\, \nabla U'(\rho_t)\rangle\, d\mu_t;$$

and similarly

$$\left.\frac{d}{ds}\right|_{s=1^-} U_\nu(\mu_t^{(s)}) = \int p(\widetilde{\rho}_t)\, Ł\widetilde{\psi}\, d\nu = \int\int \langle \nabla\widetilde{\psi},\, \nabla U'(\widetilde{\rho}_t)\rangle\, d\widetilde{\mu}_t.$$

Plugging this in (24.12) results in

$$\frac{d}{dt} W_2(\mu_t, \widetilde{\mu}_t)^2 \leq -2\,[f(t,1) - f(t,0)], \tag{24.13}$$

where

$$f(t,s) = \frac{d}{ds} U_\nu(\mu_t^{(s)}).$$

It follows from Theorem 17.15(ii) that $U_\nu(\mu_t^{(s)})$ is $K\lambda\, W_2(\mu_t, \widetilde{\mu}_t)^2$-convex on $[0,1]$ (as a function of s!!), where λ is provided by (24.7). In particular,

$$f(t,1) - f(t,0) \geq K\lambda\, W_2(\mu_t, \widetilde{\mu}_t)^2.$$

This combined with (24.13) implies the differential inequality

$$\frac{d}{dt} W_2(\mu_t, \widetilde{\mu}_t)^2 \leq -2K\lambda\, W_2(\mu_t, \widetilde{\mu}_t)^2,$$

and the desired result follows. □

Short-time behavior

A popular and important topic in the study of diffusion processes consists in establishing **regularization estimates** in short time. Typically, a certain functional used to quantify the regularity of the solution (for instance, the supremum of the unknown or some Lebesgue or Sobolev norm) is shown to be bounded like $O(t^{-\kappa})$ for some characteristic exponent κ, independently of the initial datum (or depending only on certain weak estimates on the initial datum).

Here I shall present some slightly unconventional estimates of this type. I shall only consider the case of nonnegative curvature; the general case needs more thinking.

Theorem 24.11 (Appearance of the information functionals). *Let M be a Riemannian manifold satisfying the curvature-dimension bound $\mathrm{CD}(0,\infty)$, let $\nu \in P_2(M)$, and let $U \in \mathcal{DC}_\infty$ with $U(1) = 0$. Let further $(\mu_t)_{t\geq 0}$ be a smooth solution of (24.1). Then, with the same notation as in (21.7), for all $t \geq 0$ one has*

$$t^2\, I_{U,\nu}(\mu_t) + 2t\, U_\nu(\mu_t) + W_2(\mu_t,\nu)^2 \leq W_2(\mu_0,\nu)^2.$$

In particular,

$$U_\nu(\mu_t) \leq \frac{W_2(\mu_0,\nu)^2}{2t}, \tag{24.14}$$

$$I_\nu(\mu_t) \leq \frac{W_2(\mu_0,\nu)^2}{t^2}. \tag{24.15}$$

Particular Case 24.12. In the case $U(\rho) = \rho\log\rho$, (24.14) and (24.15) become

$$H_\nu(\mu_t) \leq \frac{W_2(\mu_0,\nu)^2}{2t}, \qquad I_\nu(\mu_t) \leq \frac{W_2(\mu_0,\nu)^2}{t^2}. \tag{24.16}$$

Remark 24.13. This theorem should be thought of as an *a priori estimate*. If things are not too badly behaved, one can then remove the assumption of smoothness by a density argument, and transform (24.14), (24.15) into genuine regularization estimates.

Remark 24.14. Theorem 24.5 gave upper bounds on $U_\nu(\mu_t) - U_\nu(\nu)$ like $O(e^{-\kappa\, t})$, with a constant depending on $U_\nu(\mu_0)$. But now we can combine Theorem 24.5 with Theorem 24.11 to get an exponential decay with a constant that does not depend on $U_\nu(\mu_0)$, but only on $W_2(\mu_0, \nu)$. By an approximation argument, this will lead to results of convergence that do not need the finiteness of $U_\nu(\mu_0)$.

Remark 24.15. I would bet that the estimates in (24.16) are optimal in general (although they would deserve more thinking) as far as the dependence on μ_0 and t is concerned. On the other hand, these bounds are terrible estimates for the short-time behavior of the Kullback and Fisher informations as functions of just t. The correct scale for the Kullback information $H_\nu(\mu_t)$ is $O(\log(1/t))$, and the correct scale for the Fisher information is $O(1/t)$, as can be checked easily in the particular case when $M = \mathbb{R}^n$ and ν is the Gaussian measure.

Formal proof of Theorem 24.11. Because U_ν is displacement convex and $U_\nu(\nu) = U(1) = 0$, we have the convexity inequality

$$U_\nu(\mu) \le U_\nu(\mu) - \langle v^{(0)}, \nabla_\mu U_\nu \rangle, \tag{24.17}$$

where $v^{(0)}$ is the "velocity" at time $s = 0$ of a geodesic path $\mu^{(s)}_{0 \le s \le 1}$ joining $\mu^{(0)} = \mu$ to $\mu^{(1)} = \nu$. More explicitly,

$$\int U(\rho)\, d\nu \le - \int \rho\, \nabla U'(\rho) \cdot \nabla \psi\, d\nu, \tag{24.18}$$

where ψ is optimal in the Monge-Kantorovich problem transporting μ to ν.

On the other hand, **asdf modifier la référence** as in the proof of the Formal Theorem 24.5,

$$\frac{d^+}{dt}\, W_2(\mu_t, \nu)^2 \le 2 \left\langle \mathrm{grad}_{\mu_t} U_\nu,\, v^{(0)} \right\rangle. \tag{24.19}$$

From (24.19) and (24.18) it follows that

$$\frac{d^+}{dt}\, W_2(\mu_t, \nu)^2 \le -2\, U_\nu(\mu_t). \tag{24.20}$$

Now introduce

$$\psi(t) := a(t)\, I_{U,\nu}(\mu_t) + b(t)\, U_\nu(\mu_t) + c(t)\, W_2(\mu_t, \nu)^2,$$

where $a(t)$, $b(t)$ and $c(t)$ will be determined later.

Because of the assumption of nonnegative curvature, the quantity $I_{U,\nu}(\mu_t)$ is nonincreasing with time. (Set $K = 0$ in (24.4)(b)). Combining this with (24.20) and Theorem 24.2(ii), we get

$$\frac{d^+\psi}{dt} \le [a'(t) - b(t)] I_{U,\nu}(\mu_t) \ + [b'(t) - 2c(t)] U_\nu(\mu_t) \ + c'(t) W_2(\mu_t, \nu)^2.$$

Now choose

$$a(t) \equiv t^2, \qquad b(t) \equiv 2t, \qquad c(t) \equiv 1,$$

then ψ has to be nonincreasing as a function of t, and the conclusion follows. □

Rigorous proof of Theorem 24.11. The rigorous proof follows the same lines as the formal proof. ..... □

Remark 24.16. Inequality (24.18) is a precised version of the HWI inequality from Theorem 20.7; the HWI inequality would follow upon applying the Cauchy–Schwarz inequality.

Remark 24.17. There are stronger known regularization results in short time, for certain of the gradient flows considered in this chapter. The two main examples are

- the **Li–Yau estimates**, which give lower bounds on $\Delta \log \rho_t$, for a solution of the heat equation on a Riemanian manifold, under certain curvature-dimension conditions. For instance, if M satisfies $\mathrm{CD}(0, N)$, then

$$\Delta \log \rho_t \geq -\frac{N}{2t};$$

- the **Aronson–Bénilan estimates**, which give lower bounds on $\Delta \rho_t^{m-1}$ for solutions of the nonlinear diffusion equation $\partial_t \rho = \Delta \rho^m$ in $\mathbb{R}^n$, where $1 - 2/n < m < 1$:

$$\frac{m}{m-1} \Delta(\rho_t^{m-1}) \geq -\frac{n}{\lambda t}, \qquad \lambda = 2 - n(1-m).$$

There is an obvious similarity between these two estimates, and both can be interpreted as a lower bound on the rate of divergence of the vector field which drives particles in the gradient flow interpretation of these partial differential equations. I think it would be very interesting to have a unified proof of these inequalities, under certain geometric conditions, which would use the gradient flow interpretation of the heat and nonlinear diffusion equations, and maybe some restriction-type argument.

Open Problem 24.18. *Use a gradient flow method to extend Theorem 24.11 to the case when $K < 0$. (There are more comments about this question in the bibliographical notes.)*

Bibliographical Notes

Regularity theory for porous medium equations has been the object of many works, see in particular the synthesis works by Vázquez [**?**]. When one studies nonlinear diffusions by means of optimal transport theory, the regularity theory is the first thing to take care of; see e.g. [**?**, Appendix]. In a Riemannian context, Demange [**?**] presents many approximation arguments based on regularization, truncation, etc. in great detail. Going into these issues would have led me to considerably expand the size of this chapter; but ignoring them completely would have led to incorrect statements.

It has been known since the mid-seventies that logarithmic Sobolev inequalities yield rates of convergence to equilibrium for heat-like equations, and that these estimates are independent of the dimension. For certain problems of convergence to equilibrium involving entropy, logarithmic Sobolev inequalities are quite more convenient than spectral tools. This is especially true in infinite dimension, although logarithmic Sobolev inequalities are also very useful in finite dimension. For more information about logarithmic Sobolev inequalities, see the bibliographical notes for Chapter 21.

Around the mid-nineties, Toscani [354, 355] introduced the logarithmic Sobolev inequality in kinetic theory, where it was soon recognized to be quite convenient [139]. The links between logarithmic Sobolev inequalities and Fokker–Planck equations were reinvestigated by the kinetic theory community, see in particular [23] and the references therein. The emphasis was more on *proving* logarithmic Sobolev inequalities thanks to the study of the convergence to equilibrium for Fokker–Planck equations, than the reverse.

Soon after, it was discovered independently by Otto [290], Carrillo and Toscani [104] and Del Pino and Dolbeault [131] that the same tools could be used for nonlinear equations such as porous medium equations

$$\frac{\partial \rho}{\partial t} = \Delta \rho^m \tag{24.21}$$

in $\mathbb{R}^n$. For such equations, there is no convergence to equilibrium: the solution vanishes to infinity. But there is a well-known scaling, due to Barenblatt, under which (24.21) becomes

$$\frac{\partial \rho}{\partial t} = \Delta \rho^m + \nabla_x \cdot (\rho x). \tag{24.22}$$

Then, up to rescaling space and time, it is equivalent to understand the convergence to equilibrium for (24.22), or to understand the asymptotic behavior for (24.21), that is, how fast it approaches a certain known self-similar profile.

The extra drift term in (24.22) acts like the confinement by a quadratic potential, and this in effect is equivalent to imposing a curvature condition $\mathrm{CD}(K, \infty)$. This explains why there is an approach based on generalized logarithmic Sobolev inequalities, quite similar to the proof of Theorem 24.5.

These problems can be attacked without any particular knowledge of optimal transport. In fact, among the authors quoted before, only Otto did use optimal transport, and this was not at the level of proofs, but only at the level of intuition. Later in [292], Otto and myself gave a more direct proof of logarithmic Sobolev inequality based on the HWI inequality. The same strategy was applied again in my joint work with Carrillo and McCann [103], for more general equations involving also a (simple) nonlinear drift.

In [103] the basic equation is of the form

$$\frac{\partial \rho}{\partial t} = \sigma \Delta \rho + \nabla \cdot (\rho \nabla V) + \nabla \cdot \Big(\rho \, \nabla (\rho * \nabla W) \Big), \tag{24.23}$$

where $W = W(x - y)$ is some interaction potential on $\mathbb{R}^n$. These equations (a particular instance of McKean–Vlasov equations) appeared in the modelling of granular media [270, 44, 43], with $\sigma = 0$ or $\sigma > 0$, in particular in dimension 1. See the review paper [367] for much more informations. Similar equations also appear in the theory of self-interacting diffusion processes [38, 40, 41]. There are criteria for exponential convergence in terms of the convexity of V and W. These problems can also be set on a Riemannian manifold M (replace $W(x - y)$ by $W(x, y)$), and then Ricci curvature estimates on M come into play [338].

Demange recently studied the fast diffusion equation $\partial_t \rho = \Delta \rho^{1 - 1/N}$ on a Riemannian manifold, under a curvature-dimension condition $\mathrm{CD}(K, N)$. He used the Sobolev inequality, in the form

$$H_{N/2}(\mu) \le \frac{(N-2)(N-1)}{2K} \int \rho^{-1-\frac{2}{N}} |\nabla \rho|^2 \, d\nu \\ \le \frac{(N-2)(N-1)}{2K} (\sup \rho)^{-\frac{1}{N}} \int \rho^{1-\frac{1}{N}} |\nabla \rho|^2 \, d\nu$$

to obtain a differential inequality such as

$$\frac{dH_{N/2}(\mu_t)}{dt} \le - \left(\frac{N-2}{N-1} \right) (\sup \rho)^{-\frac{1}{N}} \, \frac{H_{N/2}(\mu_t)}{2K},$$

and deduced an estimate of the form

$$H_{N/2}(\mu_t) = O\big(e^{-(\lambda_N+\varepsilon)\,t}\big),$$

where λ_N is the presumably optimal rate that one would obtain without the $(\sup \rho)$ term, and $\varepsilon > 0$ is arbitrarily small. His estimate is slightly stronger than mine, but the asymptotic rate is the same (when measured with the same functionals).

All the methods described before apply to the study of the time asymptotics of the porous medium equation $\partial_t \rho = \Delta \rho^m$, but under the restriction $m \geq 1 - 1/N$. In that regime one can use time-rescaling and tools similar to the ones described in this chapter, to prove that the solutions become close to the so-called Barenblatt self-similar solution.

When $m < 1 - 1/N$ however, we get out of the regime where displacement convexity and related tricks can be applied. This is why it came as a bit of a sensation when Carrillo and Vázquez [105] applied the Aronson–Bénilan estimates to the problem of asymptotic behavior for fast diffusion equations with exponents m in the range $(1 - \frac{2}{N}, 1 - \frac{1}{N})$. This range of exponents is about the best one can hope, since the Barenblatt profiles do not exist for $m \leq 1 - 2/N$.

Here we see the limits of Otto's formalism: such results as the refinement of the rate of convergence of logarithmic Sobolev inequalities (Remark 24.8), or the Carrillo–Vázquez estimates, rely on inequalities of the form

$$\int p(\rho)\, \Gamma_2\big(\nabla U'(\rho)\big)\, d\nu + \int p_2(\rho)\, (LU'(\rho))^2\, d\nu \geq ....$$

in which ones takes advantage of the fact that *the same* function ρ appears in the terms $p(\rho)$ and $p_2(\rho)$ one one hand, and in the terms $\nabla U'(\rho)$ and $LU'(\rho)$ on the other hand. The technical tool might be changes of variables for the Γ_2 (as in [232]), or elementary integration by parts (as in [105]); but I don't see any interpretation of these tricks in terms of the Wasserstein space $P_2(M)$.

The story about the rates of equilibration for fast diffusion equations does not end here. At the same time as Carrillo and Vázquez obtained their main results, Denzler and McCann [**?**, 138] computed the spectral gap for the linearized fast diffusion equations in the same interval of exponents. This study showed that the rate of convergence obtained by Carrillo and Vázquez is off the value suggested by the linearized analysis by a factor 2 (except in the radially symmetric case where they obtain the optimal rate thanks to a comparison method). The connection between the nonlinear and the linearized dynamics is still unclear, although some partial results have been obtained by McCann and Slepcev [**?**]. A recent work by Cáceres and Toscani [**?**] also recovers some of the results of Denzler and McCann by means of completely different methods taking their roots in kinetic theory.

More recently, Kim and McCann [227] have derived optimal rates of convergence for the "fastest" nonlinear diffusion equations, in the range $1 - 2/N < m \leq 1 - 2/(N+2)$, by comparison methods involving Newtonian potentials. There is still ongoing research to push the rates of convergence and the range of admissible nonlinearities, in particular by Denzler, Koch, McCann and probably others.

In dimension 2, the limit case $m = 0$ corresponds to a logarithmic diffusion; it is related to geometric problems, such as the evolution of conformal surfaces or the Ricci flow [362, Chapter 8].

More general nonlinear diffusion equations of the form $\partial_t \rho = \Delta p(\rho)$ have been studied by Biler, Dolbeault and Esteban [51], Carrillo, DiFrancesco and Toscani [100, 101] in $\mathbb{R}^n$. In the latter work the rescaling procedure is recast in a more geometric and physical interpretation, in terms of temperature and projections. General nonlinear diffusion equations were also studied in a genuinely geometric setting by Demange [137] under a $\mathrm{CD}(K, N)$ curvature-dimension condition with $K > 0$.

The Csiszár–Kullback–Pinsker (CKP) inequality (24.9) was found independently by Pinsker [297], Kullback [231] and Csiszár [120]. A short and natural proof (inspired from a graduate course by Talagrand) appears in Chapter 3 of my lecture notes on measure theory and integration, available online at `www.umpa.ens-lyon.fr/~cvillani/Cours/`. See [358, 290, 63] for various generalizations of the CKP inequality.

Short-time decay estimates for the entropy and Fisher information, along the heat flow, were studied by Otto and myself around 1999 as a technical ingredient to get certain a priori estimates in a problem of hydrodynamical limits. This work was not published, and I was quite surprised to discover that Bobkov, Gentil and Ledoux [56] had found similar inequalities and applied them to get a new proof of the HWI inequality. Otto and myself published our method [293] as a comment to [56]. The method used by Otto and myself consisted in adaptating to the Wasserstein space some classical estimates about gradient flows in Hilbert spaces, that can be found in Brézis [78]. The result of Bobkov, Gentil and Ledoux is actually more general than ours, because these authors seem to have sharp constants under $\mathrm{CD}(K, \infty)$ for all values of $K \in \mathbb{R}$, while it is not clear that our method is sharp for $K \neq 0$. For $K = 0$ both methods yield exactly the same result, which was a bit of a puzzle to me. It would be interesting to clarify all this.

In relation with Remark 24.15, there is the following question which was asked to me by Guionnet (and which I am unable to answer): Given a solution (μ_t) of the heat equation $\partial_t \rho = L\rho_t$, is it true that $t\, I_\nu(\mu_t)$ converges to a finite limit, as $t \to 0$? If yes, then by De L'Hospital's rule, this is also the limit of $H_\nu(\mu_t)/\log t$ as $t \to 0$. In the particular case when $\mu_0 = f\,\nu + \sum_{k=1}^N a_k\, \delta_{x_k}$, with f smooth, it is not difficult to show that $t\, I_\nu(\mu_t)$ converges to $\sum a_k$.

Inequality (24.20) is proven rigorously in [293], under adequate regularity assumptions, for the main case of interest which is $U(r) = r \log r$.

In [102] and [15] it was investigated whether one could directly use Otto's formalism to do the proofs presented in this chapter.

The Li–Yau heat kernel estimates go back to [239]; they were refined by Davies [127], then by Bakry and Qian [35]; the latter paper is closely related to certain issues that will be addressed in the next chapter. In any case, the Bochner formula and various forms of maximum principles are the main ingredients behind these estimates. Recently, Bakry and Ledoux have derived improved forms of the Li–Yau estimates [32], and made the connection with the theory of logarithmic Sobolev inequalities.

The Aronson–Bénilan estimates were established in [25]. There is some overlap between the Aronson–Bénilan and Li–Yau bounds; together with Carrillo, I have tried without success to put both estimates in a common framework.

Apart from short-time estimates, the gradient flow approach also provides natural Hölder-1/2 estimates about the time-regularity, with values in the Wasserstein space. These estimates, that in general are essentially optimal, are discussed in [15].

25

Gradient flows III: Functional inequalities

In the preceding chapter certain functional inequalities were used to provide quantitative information about the behavior of solutions to certain partial differential equations. In the present chapter, conversely, the behavior of solutions to certain partial differential equations will help establishing certain functional inequalities.

For the kind of inequalities that will be encountered in this chapter, this principle has been explored in depth in the past two decades, starting with Bakry and Émery's heat semigroup proof of Theorem 21.2. Nowadays, one can prove this theorem by more direct means (as I did in Chapter 21); nevertheless, the heat semigroup argument is still of interest, and not only for historical reasons. It has been the basis for many generalizations, some of which are still out of reach of alternative methods.

Optimal transport appears in this game from two different perspectives. On one hand, several inequalities involving optimal transport have been proven by diffusion semigroup methods. On the other hand, optimal transport has provided a re-interpretation of the method, since several diffusion equations can be understood as gradient flows with respect to a structure induced by optimal transport. This last point of view has led to a more synthetic and geometric picture of the field; and Otto's calculus has provided a way to shortcut some intricate computations, at least from a formal point of view.

That being said, I have to admit that there are limitations to this point of view. It is true that some of the most important computations in Bakry's Γ_2 calculus can be understood in terms of optimal transport; but some other parts of the formalism, in particular those based on changes of functions, have remained inaccessible so far. Usually such manipulations are useful to treat functional inequalities involving a natural class of function whose dimension "does not match" the dimension of the curvature-dimension condition. More explicitly: It is okay to interpret in terms of optimal transport a proof involving functions in $\mathcal{DC}_\infty$ under a curvature-dimension assumption $\mathrm{CD}(K, \infty)$. Such is also the case for a proof involving functions in $\mathcal{DC}_N$ under a curvature-dimension assumption $\mathrm{CD}(K, N)$. But to get the correct constants for an inequality involving functions in $\mathcal{DC}_N$ under a curvature-dimension assumption $\mathrm{CD}(K, N')$, $N' < N$, will be much more of a problem.

In this chapter, I shall discuss three examples which can be worked out nicely. The first one is an alternative proof of Theorem 21.2, following the original argument of Bakry and Émery. The second example is a proof of the optimal Sobolev inequality (21.9) under a $\mathrm{CD}(K, N)$ condition, as recently treated by Demange. The third example is an alternative proof of Theorem 22.15, following the lines of the original proof by Otto and myself.

The proofs in this chapter will be sloppy in the sense that I shall not go into smoothness issues, or rather admit auxiliary regularity results which are not so trivial, especially in an unbounded phase space. These regularity issues are certainly the main drawback of the gradient flow approach to functional inequalities. (In fact, the first proofs based on this

strategy just ignored this problem!) More information can be found in the bibliographical notes.

In this chapter, U will be a nonlinearity belonging to a displacement convex class, and

$$U_\nu(\mu) = \int U(\rho)\, d\nu, \qquad I_{U,\nu}(\mu) = \int \rho\, |\nabla U'(\rho)|^2\, d\nu,$$

where ρ is the density of μ with respect to ν. As in Chapter 16, I shall use the notation $p(r) = rU'(r) - U(r)$ without recalling the explicit link between p and U. Further recall the following notation:

$$H_{N,\nu}(\mu) = -N \int (\rho^{1-\frac{1}{N}} - \rho)\, d\nu, \qquad I_{N,\nu}(\mu) = \left(1 - \frac{1}{N}\right)^2 \int \rho^{-1-\frac{2}{N}} |\nabla \rho|^2\, d\nu,$$

$$H_{\infty,\nu}(\mu) = H_\nu(\mu) = \int \rho \log \rho\, d\nu, \qquad I_{\infty,\nu}(\mu) = I_\nu(\mu) = \int \frac{|\nabla \rho|^2}{\rho}\, d\nu.$$

Logarithmic Sobolev inequalities revisited

Theorem 25.1 (Infinite-dimensional Sobolev inequalities from Ricci curvature). *Let M be a Riemannian manifold equipped with a reference measure ν satisfying the curvature-dimension bound $\mathrm{CD}(K,\infty)$ for some $K > 0$, and let $U \in \mathcal{DC}_\infty$. Let further $\lambda := \lim_{r\to 0} p(r)/r$. Then, for all $\mu \in P_2^{\mathrm{ac}}(M)$,*

$$U_\nu(\mu) \le \frac{I_{U,\nu}(\mu)}{2K\lambda}.$$

Particular Case 25.2 (Bakry–Émery theorem). Under the assumption $\mathrm{CD}(K,\infty)$, holds the logarithmic Sobolev inequality

$$\forall \mu \in P^{\mathrm{ac}}(M) \qquad H_\nu(\mu) \le \frac{I_\nu(\mu)}{2K}.$$

Sloppy proof of Theorem 25.1. By using Theorem 17.7(iv) and an easy approximation argument, we may assume that $U'(0) > -\infty$. Similarly, by regularizing the measure μ, we may assume that $H_\nu(\mu) < +\infty$.

Consider the gradient flow

$$\frac{\partial \rho}{\partial t} = -\nabla \cdot (\rho \nabla U'(\rho)),$$

where the initial datum ρ_0 is the density of μ. For notational simplicity, let

$$H(t) := U_\nu(\mu_t), \qquad I(t) := I_{U,\nu}(\mu_t).$$

It can be shown that $H(t)$ is continuous at $t = 0$.

From Theorem 24.5(i)(b),

$$\frac{dH(t)}{dt} = -I(t), \qquad I(t) \le I(0) e^{-2K\lambda t}.$$

These identities imply

$$H(0) = \int_0^{+\infty} I(t)\, dt \le I(0) \int_0^{+\infty} e^{-2K\lambda t}\, dt = \frac{I(0)}{2K\lambda}.$$

□

Sobolev inequalities revisited

In this section I shall reformulate the argument recently used by Demange to recover the optimal Sobolev inequalities (21.10) via a gradient flow approach.

Theorem 25.3 (generalized Sobolev inequalities under Ricci curvature bounds). *Let M be a Riemannian manifold equipped with a reference measure ν, satisfying a curvature-dimension bound $\mathrm{CD}(K,N)$ for some $K > 0$, $N \in [1,\infty)$. Let $U \in \mathcal{DC}_N$, and let $A : \mathbb{R}_+ \to \mathbb{R}$ be such that $A(0) = A(1) = 0$ and $A''(r) = r^{-\frac{1}{N}} U''(r)$. Then, for any probability density ρ on M,*

$$\int_M A(\rho)\, d\nu \le \frac{1}{2K\lambda} \int_M \rho\, |\nabla U'(\rho)|^2\, d\nu, \tag{25.1}$$

where

$$\lambda = \lim_{r \downarrow 0} \frac{p(r)}{r^{1-\frac{1}{N}}}.$$

Particular Case 25.4 (Sobolev inequalities). Choose

$$U(r) = U_N(r) = -N(r^{1-\frac{1}{N}} - r), \qquad A(r) = -\frac{N(N-1)}{2(N-2)}\,(r^{1-\frac{2}{N}} - r);$$

then (25.1) reads

$$H_{\frac{N}{2},\nu}(\mu) \le \frac{1}{2K} \left(\frac{N-2}{N-1}\right) I_{N,\nu}(\mu),$$

which can also be rewritten in the form of (21.10) or (21.9).

Sloppy proof of Theorem 25.3. By density, we may assume that the density ρ_0 of μ is smooth; we may also assume that A and U are smooth. Let $(\rho_t)_{t \ge 0}$ be the solution of the gradient flow equation

$$\frac{\partial \rho}{\partial t} = \nabla \cdot (\rho \nabla U'(\rho)), \tag{25.2}$$

and as usual $\mu_t = \rho_t\, \nu$.

By Theorem 24.2(iii),

$$\frac{d}{dt} I_{U,\nu}(\mu_t) \le -2K\lambda \int_M \rho_t^{1-\frac{1}{N}} |\nabla U'(\rho_t)|^2\, d\nu. \tag{25.3}$$

On the other hand, from the assumption $A''(r) = r^{-\frac{1}{N}} U''(r)$

$$\nabla A'(\rho) = \rho^{-\frac{1}{N}}\, \nabla U'(\rho).$$

So Theorem 24.2(i) implies

$$\begin{aligned} \frac{d}{dt} \int A(\rho_t)\, d\nu &= -\int_M \rho_t \nabla A'(\rho_t) \cdot \nabla U'(\rho_t)\, d\nu \\ &= -\int_M \rho_t^{1-\frac{1}{N}} \left|\nabla U'(\rho_t)\right|^2 d\nu. \end{aligned}$$

By combining this with (25.3), we obtain

$$-\frac{d}{dt} A_\nu(\mu_t) \le -\left(\frac{1}{2K\lambda}\right) \frac{d}{dt}\, I_{U,\nu}(\mu_t). \tag{25.4}$$

As $t \to \infty$, μ_t converges to ν (Theorem 24.5); so both $A_\nu(\mu_t)$ and $I_{U,\nu}(\mu_t)$ converge to 0. Then one can integrate both sides of (25.4) from $t = 0$ to $t = \infty$, and recover

$$A_\nu(\mu_0) \le \left(\frac{1}{2K\lambda}\right) I_{U,\nu}(\mu_0).$$

This concludes the proof. □

From log Sobolev to Talagrand, revisited

This section is devoted to an alternative proof of Theorem 22.15, due to Otto and myself.

Formal Theorem 25.5 (From Sobolev-type inequalities to concentration-type inequalities). *Let $U \in \mathcal{DC}_\infty$. Let M be a Riemannian manifold with Ricci curvature uniformly bounded below, equipped with a reference measure $\nu \in P_2^{\rm ac}(M)$. Assume that for any $\mu \in P_2^{\rm ac}(M)$, holds the inequality*

$$U_\nu(\mu) - U_\nu(\nu) \le \frac{1}{2K\lambda}\, I_{U,\nu}(\mu).$$

Then, for any $\mu \in P_2^{\rm ac}(M)$, also holds the inequality

$$W_2^2(\mu,\nu) \le \frac{2\,U_\nu(\mu)}{K\lambda}.$$

Particular Case 25.6 (From Log Sobolev to Talagrand). If the reference measure ν on M satisfies a logarithmic Sobolev inequality with constant K, then it also satisfies a Talagrand inequality with constant K:

$$\forall \mu \in P^{\rm ac}(M), \qquad H_\nu(\mu) \le \frac{I_\nu(\mu)}{2K}$$

implies

$$\forall \mu \in P^{\rm ac}(M), \qquad W_2(\mu,\nu) \le \sqrt{\frac{2\,H_\nu(\mu)}{K}}. \tag{25.5}$$

Sloppy proof of Theorem 25.5. By a density argument, we may assume that μ has a smooth density μ_0, and let $(\mu_t)_{t\ge 0}$ evolve according to the gradient flow (25.2). By Theorem 24.2(ii),

$$\frac{d}{dt} U_\nu(\mu_t) = -I_{U,\nu}(\mu_t).$$

asdf RÉÉCRIRE CE MORCEAU DE LA PREUVE
It follows that, for any $t > 0$,

$$\frac{d^+}{dt} W_2(\mu_0,\mu_t) \le \left(\frac{1}{2W_2(\mu_0,\mu_t)}\right) \frac{d^+}{dt} W_2(\mu_0,\mu_t)^2 \le \sqrt{I_{U,\nu}(\mu_t)}. \tag{25.6}$$

On the other hand, by assumption,

$$\sqrt{I_{U,\nu}(\mu_t)} \le \frac{I_{U,\nu}(\mu_t)}{\sqrt{2K\lambda\, U_\nu(\mu_t)}} = -\frac{d}{dt}\sqrt{\frac{K\lambda\, U_\nu(\mu_t)}{2}}. \tag{25.7}$$

From (25.6) and (25.7),

$$\frac{d^+}{dt} W_2(\mu_0, \mu_t) \le -\frac{d}{dt} \sqrt{\frac{K\lambda\, U_\nu(\mu_t)}{2}}.$$

Stated otherwise: If

$$\psi(t) := W_2(\mu_0, \mu_t) + \sqrt{\frac{K\lambda\, U_\nu(\mu_t)}{2}},$$

then $d^+\psi/dt \le 0$, i.e. ψ is nonincreasing as a function of t, and so

$$\lim_{t\to\infty} \psi(t) \le \psi(0). \tag{25.8}$$

The assumptions imply that $U_\nu(\mu_t) - U_\nu(\nu)$ converges to 0 (exponentially fast), so μ_t converges to ν. Admitting that we can pass to the limit in W_2 **asdf clarifier !!**, we see that $\psi(+\infty) = W_2(\mu_0, \nu)$, and $\psi(0) = \sqrt{(K\lambda U_\nu(\mu_0))/2}$. So (25.8) coincides with (25.5). □

Appendix: Comparison of proofs

The proofs in the present chapter were based on gradient flows, while proofs in Chapters 21 and 22 were more directly based on displacement interpolation. How do these two strategies compare?

From a formal point of view, they are not so different that one may think. Take the case of the heat equation,

$$\frac{\partial \rho}{\partial t} = \Delta \rho,$$

or equivalently

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \Big(\rho \nabla(-\log \rho)\Big) = 0.$$

The evolution of ρ is determined by the "vector field" $\rho \to (-\log \rho)$, in the space of probability densities. Rescale time and the vector field itself as follows:

$$\varphi_\varepsilon(t, x) = -\varepsilon \log \rho\left(\frac{\varepsilon t}{2},\, x\right).$$

Then φ_ε satisfies the equation

$$\frac{\partial \varphi_\varepsilon}{\partial t} + \frac{|\nabla \varphi_\varepsilon|^2}{2} = \frac{\varepsilon}{2} \Delta \varphi_\varepsilon.$$

Passing to the limit as $\varepsilon \to 0$, one gets, at least formally, the Hamilton–Jacobi equation

$$\frac{\partial \varphi}{\partial t} + \frac{|\nabla \varphi|^2}{2} = 0,$$

which is in some sense the equation driving displacement interpolation.

There is a general principle here: After suitable rescaling, the velocity field associated with a gradient flow resembles the velocity field of a geodesic flow. Here might be a possible way to see this. Take an arbitrary smooth function U, and consider the evolution

$$\dot{x}(t) = -\nabla U(x(t)).$$

Turn to Eulerian formalism, consider the associated vector field v defined by

$$\frac{d}{dt} X(t, x_0) = -\nabla U(X(t, x_0)) =: -v\big(t, X(t, x_0)\big),$$

and rescale by

$$v_\varepsilon(t, x_0) = \varepsilon\, v\big(\varepsilon t,\, X(\varepsilon t, x_0)\big).$$

then one can check that, as $\varepsilon \to 0$,

$$\nabla_{x_0} v_\varepsilon(t, x_0) \simeq \varepsilon \nabla^2 U(x_0).$$

It follows by an explicit calculation that

$$\frac{\partial v_\varepsilon}{\partial t} + v_\varepsilon \cdot \nabla v_\varepsilon \simeq 0.$$

So as $\varepsilon \to 0$, $v_\varepsilon(t, x)$ should asymptotically satisfy the equation of a geodesic vector field (pressureless Euler equation).

There is probably more to say on the subject, but whatever the interpretation, the Hamilton–Jacobi equations can always be squeezed out of the gradient flow equations after some suitable rescaling. Thus we may expect the gradient flow strategy to be more precise than the displacement convexity strategy. This is also what suggests the use of Otto's calculus: proofs based on the gradient flow approach need a control of $\mathrm{Hess}\, U$ in the direction ∇U, while proofs based on the displacement convexity approach need a control of $\mathrm{Hess}\, U$ in all directions. This might explain why there is at present no displacement convexity analogue of Demange's gradient flow proof of the Sobolev inequality (so far only weaker inequalities with nonsharp constants have been obtained).

On the other hand, proofs based on displacement convexity are usually quite simpler, and quite more robust than proofs based on gradient flows: no issues about the regularity of the semigroup, no subtle interplay between the Hessian of the functional and the "direction of evolution"...

To summarize: Proofs based on displacement convexity are simpler and conceptually easier, but proofs based on gradient flows seem to apply in more situations, and may lead to sharper results.

In the end we can put some of the main functional inequalities discussed in these notes in a nice array. Below, "LSI" stands for "Logarithmic Sobolev inequality"; "T" for "Talagrand inequality"; and "Sob_2" for the Sobolev inequality with exponent 2. So $\mathrm{LSI}(K)$, $\mathrm{T}(K)$, $\mathrm{HWI}(K)$ and $\mathrm{Sob}_2(K, N)$ respectively stand for (21.4), (**??**) (with $p = 2$), (20.10) and (21.9).

Theorem	Gradient flow proof	Displacement convexity proof
$\mathrm{CD}(K, \infty) \Rightarrow \mathrm{LSI}(K)$	Bakry–Émery	Otto–Villani
$\mathrm{LSI}(K) \Rightarrow \mathrm{T}(K)$	Otto–Villani	Bobkov–Gentil–Ledoux
$\mathrm{CD}(K, \infty) \Rightarrow \mathrm{HWI}(K)$	Bobkov–Gentil–Ledoux	Otto–Villani
$\mathrm{CD}(K, N) \Rightarrow \mathrm{Sob}_2(K, N)$	Demange	??

Bibliographical Notes

The observation that the Fisher information I_ν is the time-derivative of the entropy functional $-H_\nu$ along the heat semigroup seems to first appear in a famous paper by Stam [333] at the end of the fifties, in the case $M = \mathbb{R}$ (equipped with the Lebesgue measure). Stam gives credit to de Bruijn for that remark. He exploited it to prove an inequality which can be recast (after simple changes of functions) as the usual logarithmic Sobolev inequality, found fifteen years later by Gross [198]. Stam's inequality reads $N\, I \geq 1$, where I is the Fisher information, and N is the "power entropy". (In dimension n, this inequality should be replaced by $N\, I \geq n$.) The main difference between these inequalities

is that Stam's one is expressed in terms of the Lebesgue reference measure, while Gross's one is expressed in terms of the Gaussian reference measure. Although Stam is famous for his information-theoretical inequalities, it is only at the beginning of the nineties that specialists identified a version of the logarithmic Sobolev inequality in his work.

Stam's argument is not completely rigorous because of regularity issues, but can be repaired; see for instance [353, 96].

The proof of Theorem 25.1 in this chapter follows the strategy by Bakry and Émery, who were only interested in the Particular Case 25.2. Bakry and Émery used a set of calculus rules which has sometimes been dubbed as the "Γ_2 calculus". They were not very careful about regularity issues, and for that reason I think that the original proof cannot be considered as rigorous (in particular for noncompact manifolds, in which regularity issues are not so innocent, even if the curvature-dimension condition prevents the blow-up of the heat semigroup). However, recently Demange [137] carried out complete proofs for much more delicate situations, so there is no reason to doubt that the Bakry–Émery argument can be made completely rigorous. Also, when the manifold is $\mathbb{R}^n$ equipped with a reference density e^{-V}, the Bakry–Émery proof was carefully rewritten by Arnold, Markowich, Toscani and Unterreiter [23], in the language of partial differential equations. This paper was the sequel of a simpler paper by Toscani [355] considering the particular case of the Gaussian measure.

The Bakry–Émery strategy was applied independently by Otto [290] and by Carrillo and Toscani [104] to study the asymptotic behavior of porous medium equations. Since then, many authors have applied it to various classes of nonlinear equations, see e.g. [103, 105].

The interpretation of the Bakry–Émery proof as a gradient flow argument was developed in my paper with Otto [292]. This interpretation was of much help when we considered the more complicated nonlinear situation in [103].

Demange [137] did not only treat the inequality (21.10), but also the whole family (21.8). A disturbing remark is that for many members of this family, there is no uniqueness of the gradient flow that one can use in the proof. He also discusses other criteria than $U \in \mathcal{DC}_N$, allowing for finer results if, say, $U \in \mathcal{DC}_N$ but the curvature-dimension bound is $\mathrm{CD}(K, N')$ for some $N' < N$; at this point he uses formulas of change of variables for Γ_2 operators. He found a mysterious structure condition on the nonlinearity U, which in many cases leads to finer results than the $\mathcal{DC}_N$ condition:

$$rq'(r) + q(r) \geq \frac{9N}{4(N+2)}\, q^2(r), \qquad q(r) = \frac{rU''(r)}{U'(r)} + \frac{1}{N}. \tag{25.9}$$

Demange worked on arbitrary noncompact manifolds by using a careful truncation procedure. He restricts the equation to bounded open subsets and imposes Dirichlet boundary conditions. (Neumann's boundary condition would be more natural, for instance because it preserves the mass; but the Dirichlet boundary conditions have the major technical advantage to be associated with a monotonicity principle.) All of Demange's results still seem to be out of reach of more direct methods based on displacement interpolation.

The proof of Theorem 25.5 was implemented in my joint work with Otto [292]. The proof there is (hopefully!) complete, but we only considered the Particular Case 25.6 (certainly the most important). We imposed the Ricci curvature of the manifold to be bounded below, so as to avoid any blow-up of the heat semigroup. Maybe one can still make the proof work without that lower bound assumption, by truncating the logarithmic Sobolev inequality and the Talagrand inequality, and then work in an arbitrarily large bounded open subset of the manifold, imposing Neumann boundary conditions. In any case, to treat noncompact manifolds without lower bounds on the curvature, it is certainly easier to use the proof of Theorem 22.15.

The observation that the Hamilton–Jacobi equation can be obtained from the heat equation after proper rescaling is quite old, and it is now a classical exercise in the theory of viscosity solutions, see e.g. Barles [37]. Bobkov, Gentil and Ledoux [56] observed that this could constitute a bridge between the two main existing strategies for logarithmic Sobolev inequalities. Links with the theory of large deviations have been investigated in [163].

As for the final array, the corresponding papers are those of Bakry–Émery [31], Otto–Villani [292], Bobkov–Gentil–Ledoux [56], Demange [136, 137].

Part III

Synthetic treatment of Ricci curvature

The last part of these notes is devoted to a recent direction of research which was mainly explored by Lott, Sturm and myself from 2004 on.

In Chapter 17 it was proven that lower Ricci curvature bounds influence displacement convexity properties of certain classes of functionals; but also that these properties **characterize** lower Ricci curvature bounds. So we may "transform the theorem into a definition" and express the property "Ricci curvature is bounded below by K" in terms of certain displacement convexity properties. This approach is **synthetic**, in the sense that it does not rely on analytic computations (of the Ricci tensor...), but rather on the properties of certain objects which play an important role in some geometric arguments.

This point of view has the advantage to apply to nonsmooth spaces, just as lower (or upper) sectional curvature bounds can be defined in nonsmooth metric spaces by Alexandrov's method. An important difference however is that the notion of generalized Ricci curvature will be defined not only in terms of distances, but also in terms of reference measures. So the basic object will not be a metric space, but a **metric-measure space**, that is a metric space equipped with a reference measure.

Chapters 26 and 27 are preparatory. In Chapter 26 I shall try to convey in some detail the meaning of the word "synthetic", with a simple illustration about convex functions; then Chapter 27 will be devoted to some reminders about the convergence of metric-measure spaces.

The next two chapters constitute the core of this part. In Chapter 28 I will consider optimal transport in possibly nonsmooth spaces, and establish various properties of *stability* of optimal transport under convergence of metric-measure spaces. Then in Chapter 29 I shall present a synthetic definition of the curvature-dimension condition $\mathrm{CD}(K, N)$ in a nonsmooth context, and prove that it too is stable. Here is a geometric consequence of these results, that can be stated without any reference to optimal transport: If a Riemannian manifold is the limit of a sequence of $\mathrm{CD}(K, N)$ Riemannian manifolds, then it, too, satisfies $\mathrm{CD}(K, N)$.

The last chapter will present a state of the art about the qualitative geometric and analytic properties enjoyed by metric-measure spaces satisfying curvature-dimension conditions.

26

Analytic and synthetic points of view

The present chapter is devoted to a simple pedagogical illustration of the opposition between the "analytic" and "synthetic" points of view. Consider the following two definitions for convexity on $\mathbb{R}^n$:

(i) A convex function is a function φ which is twice continuously differentiable, and whose Hessian $\nabla_x^2 \varphi$ is nonnegative at each $x \in \mathbb{R}^n$;

(ii) A convex function is a function φ such that for all $x, y \in \mathbb{R}^n$, and $\lambda \in [0,1]$,

$$\varphi\big((1-\lambda)x + \lambda y\big) \leq (1-\lambda)\,\varphi(x) + \lambda\,\varphi(y).$$

How can we compare these two definitions?

1) When applied to C^2 functions, both definitions coincide, but the second one is obviously *more general*. Not only is it expressed without any reference to second differentiability, but there are examples, such as $\varphi(x) = |x|$, which satisfy (ii) but not (i). So Definition (ii) really is an extension of Definition (i).

2) Definition (ii) is *more stable* than Definition (i). Here is what I mean by that: Take a sequence $(\varphi_k)_{k\in\mathbb{N}}$ of convex functions, converging to some other function φ, how do I know that φ is convex? To pass to the limit in Definition (i), I would need the convergence to be very strong, say in $C^2(\mathbb{R}^n)$. (Let's forget here about the notion of distributional convergence, which would solve the problem but is much less elementary.) On the other hand, I can pass to the limit in Definition (ii) assuming only, say, pointwise convergence. So Definition (ii) is much easier to "pass to the limit in" — even if the limit is known to be smooth.

3) Definition (ii) is also a better *starting point* for studying properties of convex functions. In this set of notes, most of the time, when I used some convexity, it was via (ii), not (i).

4) On the other hand, if I give you a particular function (by its explicit analytic expression, for instance), and ask you whether it is convex, it will probably be a nightmare to check Definition (ii) directly, while Definition (i) might be workable: You just need to *compute* the second derivative and check its sign. Probably this is the method that will work most easily for the huge majority of "potentially convex" functions that you will meet in your life, if you don't have any extra information on them (like they are the limit of some family of functions...).

5) Definition (i) is naturally *local*, while Definition (ii) is global (and probably this is related to the fact that it is so difficult to check). In particular, Definition (i) involves an

object (the second derivative) which can be used to quantify the "strength of convexity" *at each point*). Of course, I can always define a convex function as a function which satisfies (ii) *locally*, i.e. when x and y stay in the neighborhood of any given point; but then locality does not enter in such a simple way as in (i), and the issue immediately arises whether a function which satisfies (ii) locally, also satisfies (ii) globally.

In the above discussion, Definition (i) can be thought of as **analytic** (it is based on the computation of certain objects), while Definition (ii) is **synthetic** (it is based on certain qualitative properties which are the basis for proofs). Observations 1–5 above are in some sense typical: synthetic definitions ought to be more general and more stable, and they should be usable directly to prove interesting results; on the other hand, they may be difficult to check in practice, and they are usually less precise (and less "local") than analytic definitions.

In classical Euclidean geometry, the analytic approach consists in introducing Cartesian coordinates and making computations with equations of lines and circles, sines and cosines, etc. The synthetic approach, on the other hand, is more or less the one that was used already by ancient Greeks (and which is still taught, or at least should be taught, to our kids, for developing the intuition of proof-making): it is not based on computations, but on axioms *à la* Euclid, qualitative properties of lines, angles, circles and triangles, construction of auxiliary points, etc. The analytic approach is conceptually simple, but sometimes leads to very horrible computations; the synthetic approach is often lighter, but requires a better intuition and clever elementary arguments. It is also usually (but this is of course a matter of taste) more elegant.

In "Riemannian" geometry, curvature is usually defined via a purely analytic approach: From the Riemannian scalar product one can compute several functions which are called sectional curvature, Ricci curvature, scalar curvature, etc. For instance, for any $x \in M$, the sectional curvature at point x is a *function* which associates to each 2-dimensional plane $P \subset T_x M$ a number $\sigma_x(P)$, for which there is an explicit expression in terms of a basis of P, and a certain combination of derivatives of the metric at x. Intuitively, $\sigma_x(P)$ measures the speed of convergence of geodesics that start at x, with velocities spanning the plane P. A lot of geometric information can be retrieved from the bounds on the sectional curvature. Then a space is said to have nonnegative sectional curvature if $\sigma_x(P)$ is nonnegative, for all P and for all x.

However, there is also a famous synthetic point of view on sectional curvature, due to Alexandrov. In Alexandrov's approach one does not try to define what is the curvature, but what it means to have nonnegative curvature: By definition, a length space $(\mathcal{X}, d)$ is said to have Alexandrov curvature bounded below by K if its triangles are no more "skinny" than reference triangles drawn on the model space $\mathbb{R}^2$. More precisely: If xyz is a triangle in $\mathcal{X}$ and $x_0y_0z_0$ is a triangle drawn on $\mathbb{R}^2$ with $d(x_0, y_0) = d(x, y)$, $d(y_0, z_0) = d(y, z)$, $d(z_0, x_0) = d(z, x)$, and x' is a midpoint between y and z, x'_0 a midpoint between y_0 and z_0, then one should have $d(x_0, x'_0) \leq d(x, x')$.

There is an excellent analogy with the previous discussion for convex functions. The Alexandrov definition is equivalent to the analytic one in case it is applied to a smooth Riemannian manifold; but it is more general, since it applies for instance to a cone (say, the two-dimensional cone embeddeed in $\mathbb{R}^3$, constructed over a circular basis). It is also more stable; in particular, it passes to the limit under **Gromov–Hausdorff convergence**, a notion that will be described in the sequel. It can still be used as the starting point for many properties involving sectional curvature. On the other hand, it is in general

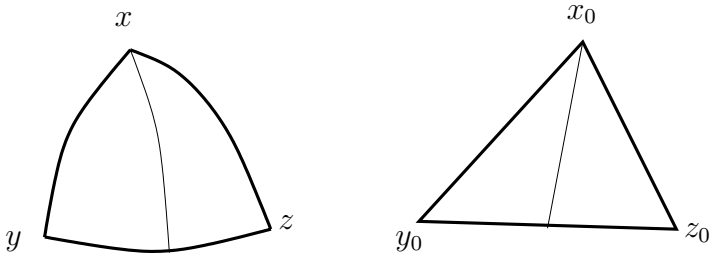
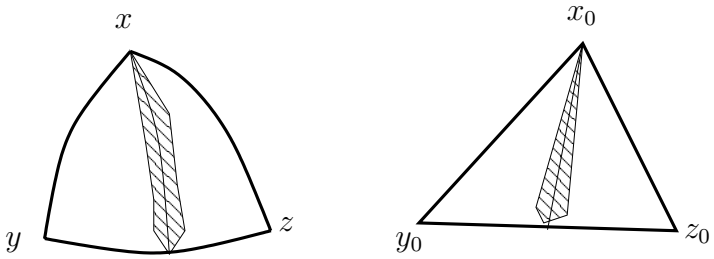


Fig. 26.1. The triangle on the left is drawn in $\mathcal{X}$, the triangle on the right is drawn on the model space $\mathbb{R}^2$; the lengths of their edges are the same. The thin geodesic lines go through the apex to the middle of the basis; the one on the left is longer than the one on the right. In that sense the triangle on the left is **fatter** than the triangle on the right. If all triangles in $\mathcal{X}$ look like this, then $\mathcal{X}$ has nonnegative curvature. (Think of a triangle as the belly of some individual, the belt being the basis, and the neck being the apex; then the line going from apex to middle of the basis is of course the tie. The fatter the individual, the longer his tie should be.)



difficult to check directly, and there is no associated notion of *curvature* (when one says "Alexandrov space of nonnegative curvature", the words "nonnegative" and "curvature" do not make sense independently of each other).

Still there is a generalization of what it means to have curvature bounded below by $K \in \mathbb{R}$, where K is an arbitrary real number, not necessarily 0. It is obtained by replacing the model space $\mathbb{R}^2$ by the model space with constant curvature K, that is

- the sphere $S^2(1/\sqrt{K})$ with radius $R = 1/\sqrt{K}$, if $K > 0$;

- the plane $\mathbb{R}^2$, if $K = 0$;

- the hyperbolic space $\mathbb{H}(1/\sqrt{|K|})$ with "hyperbolic radius" $R = 1/\sqrt{|K|}$, if $K < 0$; this is the half-plane $\mathbb{R} \times (0, +\infty)$, equipped with the metric $g_{(x,y)}(dx\, dy) = (dx^2 + dy^2)/(|K|y)$.

Length spaces which satisfy these inequalities are called Alexandrov spaces with curvature bounded below by K; all the remarks which were made above in the case $K = 0$ apply in this more general case. There is also a symmetric notion of Alexandrov spaces with curvature bounded above, obtained by just reversing the inequalities.

This generalized notion of sectional curvature bounds has been explored by many authors, and quite strong results have been obtained about the geometric and analytic implications of such bounds. But up to recently this was the only kind of curvature bounds for which a synthetic approach was available. In particular, it has been an open problem for quite some time to find a synthetic treatment of lower Ricci curvature bounds. The thesis developed in the sequel of these notes, is that optimal transport provides a solution to this problem. Of course, this might not be the only solution, but so far it looks like the only acceptable one that is available.

Bibliographical Notes

In close relation to the topics discussed in this chapter, there is an illuminating course by Gromov [195], which I strongly recommend to the reader who wants to learn about the meaning of curvature.

There are several good sources for Alexandrov spaces, in particular the book by Burago, Burago and Ivanov [81]. Such spaces are also called CAT spaces, in reference to Cartan, Alexandrov and Toponogov. But the terminology of CAT space is often restricted to Alexandrov spaces with *upper* sectional bounds. So a CAT(K) space typically means an Alexandrov space with "sectional curvature bounded above by K".

The open problem to find a satisfactory synthetic treatment of Ricci curvature bounds was discussed in the above-mentioned book by Gromov [195, pp. 84–85], and more recently in Cheeger and Colding [110, Appendix 2].

References about recent developments related to optimal transport will be given in the sequel.

27

Convergence of metric-measure spaces

The central question in this chapter is the following: What does it mean to say that a metric-measure space $(\mathcal{X}, d_{\mathcal{X}}, \nu_{\mathcal{X}})$ is "close" to another metric-measure space $(\mathcal{Y}, d_{\mathcal{Y}}, \nu_{\mathcal{Y}})$? We would like to have an answer that is as "intrinsic" as possible, in the sense that it should depend only on the metric-measure properties of $\mathcal{X}$ and $\mathcal{Y}$.

So as not to inflate the volume of these notes too much, I shall admit many proofs when they can be found in accessible references, and prefer to insist on the main stream of ideas.

Hausdorff topology

There is a well-established notion of distance between compact sets of a given metric space, namely the **Hausdorff distance**. If $\mathcal{X}$ and $\mathcal{Y}$ are two compact subsets of a metric space $(\mathcal{Z}, d)$, their Hausdorff distance is

$$d_H(\mathcal{X}, \mathcal{Y}) = \max\Big(\sup_{x \in \mathcal{X}} d(x, \mathcal{Y}),\ \sup_{y \in \mathcal{Y}} d(y, \mathcal{X})\Big),$$

where as usual $d(a, B) = \inf\{d(a, b);\ b \in B\}$ is the distance from the point a to the set B. Here the choice of notation is not innocent: Think of $\mathcal{X}$ and $\mathcal{Y}$ not just as subsets, but rather as metric subspaces of $\mathcal{Z}$.

The statement "$d_H(\mathcal{X}, \mathcal{Y}) \leq r$" can be recast informally as follows: "If we *inflate* (enlarge) $\mathcal{Y}$ by a distance r, then the resulting set covers $\mathcal{X}$; and conversely if we inflate $\mathcal{X}$ by a distance r, the resulting set covers $\mathcal{Y}$."

Fig. 27.1. In solid lines, the borders of the two sets A and B; in dashed lines, the borders of their enlargements.

The Hausdorff distance can be thought of as a set-theoretical analogue of the Prokhorov distance between probability measures (of course, historically the former came first). This will become more apparent if I rewrite the Hausdorff distance as

$$d_H(A,B) = \inf\Big\{r>0; \quad A \subset B^{r]} \ \text{ and } \ B \subset A^{r]}\Big\},$$

and rewrite the Prokhorov distance as

$$d_P(\mu,\nu) = \inf\Big\{r>0; \quad \text{for all closed } C, \quad \mu[C] \le \nu[C^{r]}] + r \ \text{ and } \nu[C] \le \mu[C^{r]}] + r\Big\},$$

where $C^{r]}$ is the set of all points whose distance to C is no more than r, i.e. the union of all *closed* balls $B_{r]}(x)$, $x \in C$.

The analogy between the two notions goes further: While the Prokhorov distance can be defined in terms of couplings, the Hausdorff distance can be defined in terms of **correspondences**. By definition, a correspondence (or relation) between two sets $\mathcal{X}$ and $\mathcal{Y}$ is a subset $\mathcal{R}$ of $\mathcal{X} \times \mathcal{Y}$; if $(x,y) \in \mathcal{R}$, then x and y are said to be in correspondence, and it is required that each $x \in \mathcal{X}$ should be in correspondence with at least one y, and each $y \in \mathcal{Y}$ should be in correspondence with at least one x.

Then we have the two very similar formulas:

$$\begin{cases} d_P(\mu,\nu) = \inf\Big\{r>0; \quad \exists \text{ coupling } (X,Y) \text{ of } (\mu,\nu); \quad \mathbb{P}\left[d(X,Y) > r\right] \le r\Big\}; \\ \\ d_H(\mu,\nu) = \inf\Big\{r>0; \quad \exists \text{ correspondence } \mathcal{R} \text{ in } \mathcal{X}\times\mathcal{Y}; \quad \forall (x,y) \in \mathcal{R}, \quad d(x,y) \le r\Big\}. \end{cases}$$

Moreover, it is easy to guess an "optimal" correspondence: Just define

$$(x,y) \in \mathcal{R} \quad \Longleftrightarrow \quad \big[\, d(x,y) = d(x,\mathcal{Y}) \text{ or } d(y,x) = d(y,\mathcal{X}) \,\big].$$

Then each $(x,y) \in \mathcal{R}$ satisfies $d(x,y) \le d_H(\mathcal{X},\mathcal{Y})$, with equality for at least one couple. (Indeed, the maximum in the definition of the Hausdorff distance is obviously achieved.)

Just like their probabilistic counterparts, correspondences can be **glued** together: if $\mathcal{R}_{12}$ is a correspondence between $\mathcal{X}_1$ and $\mathcal{X}_2$, and $\mathcal{R}_{23}$ is a correspondence between $\mathcal{X}_2$ and $\mathcal{X}_3$, then one may define a correpondence $\mathcal{R}_{13} = \mathcal{R}_{23} \circ \mathcal{R}_{12}$ between $\mathcal{X}_1$ and $\mathcal{X}_3$ by

$$(x_1,x_3) \in \mathcal{R}_{13} \quad \Longleftrightarrow \quad \big[\, \exists x_2 \in \mathcal{X}_2; \ (x_1,x_2) \in \mathcal{R}_{12} \text{ and } (x_2,x_3) \in \mathcal{R}_{23} \,\big].$$

The Hausdorff distance really is a distance! Indeed,

(i) it is obviously symmetric ($d_H(\mathcal{X},\mathcal{Y}) = d_H(\mathcal{Y},\mathcal{X})$);

(ii) because it is defined on compact (hence bounded) sets, the infimum in the definition is a nonnegative *finite* number;

(iii) if $d_H(\mathcal{X},\mathcal{Y}) = 0$, this means that any $x \in \mathcal{X}$ satisfies $d(x,\mathcal{Y}) \le \varepsilon$, for any $\varepsilon > 0$; so $d(x,\mathcal{Y}) = 0$, and since $\mathcal{Y}$ is assumed to be compact (hence closed), this implies that $\mathcal{X} = \mathcal{Y}$;

(iv) if $\mathcal{X}_1$, $\mathcal{X}_2$ and $\mathcal{X}_3$ are given, introduce optimal correspondences $\mathcal{R}_{12}$ and $\mathcal{R}_{23}$ in the correspondence representation of the Hausdorff measure; then the composed representation $\mathcal{R}_{13} = \mathcal{R}_{23} \circ \mathcal{R}_{12}$ is such that any $(x_1,x_3) \in \mathcal{R}_{13}$ satisfies $d(x_1,x_3) \le d(x_1,x_2) + d(x_2,x_3) \le d_H(\mathcal{X}_1,\mathcal{X}_2) + d_H(\mathcal{X}_2,\mathcal{X}_3)$ for some x_2.

Then one may define the metric space $\mathcal{H}(\mathcal{Z})$ as the space of all compact subsets of $\mathcal{Z}$, equipped with the Hausdorff distance. There is a nice statement that if $\mathcal{Z}$ is compact then $\mathcal{H}(\mathcal{Z})$ is also a compact metric space.

So far everything is quite simple, but soon it will become a bit more complicated, which was a good reason to go slowly.

The Gromov–Hausdorff distance

The Hausdorff distance only compares subsets of a given underlying space. But how can we compare different metric spaces with possibly nothing in common? First one would like to say that two spaces which are *isometric* really are the same. Recall the definition of an isometry: If $(\mathcal{X}, d)$ and $(\mathcal{X}', d')$ are two metric spaces, a map $f : \mathcal{X} \to \mathcal{X}'$ is called an isometry if

(a) it preserves distances: for any pair of points $x, y \in \mathcal{X}$, $d'(f(x), f(y)) = d(x, y)$;

(b) it is surjective: for any $x' \in \mathcal{X}'$ there is $x \in \mathcal{X}$ with $f(x) = x'$.

An isometry is automatically injective, so it has to be a bijection, and its inverse f^{-1} is also an isometry. Two metric spaces are said to be **isometric** if there exists an isometry between them. If two spaces $\mathcal{X}$ and $\mathcal{X}'$ are isometric, then any statement which holds true on $\mathcal{X}$ and can be expressed in terms of just the distance, will automatically be "transported" to $\mathcal{X}'$ by the isometry.

This motivates the desire to work with *isometry classes*, rather than metric space. By definition, an isometry class $\overline{\mathcal{X}}$ is the set of all metric spaces which are isometric to some given space $\mathcal{X}$. Instead of "isometry class", I shall often write "abstract metric space". All the spaces in a given isometry class have the same topological properties, so it makes sense to say of an abstract metric space that it is compact, or complete, etc.

This looks good, but a bit frightening: There are so many metric spaces around that the concept of abstract metric space seems to be ill-posed from the set-theoretical point of view (just like there is no "set of all sets"). However, things becomes much more friendly when one realizes that any compact metric space, being *separable*, is isometric to the completion of $\mathbb{N}$ for a suitable metric. (To see this, introduce a dense sequence (x_k) in your favorite space $\mathcal{X}$, and define $d(k, \ell) = d_{\mathcal{X}}(x_k, x_\ell)$.) Then we might think of an isometry class as a subset of the set of all distances on $\mathbb{N}$; this is still huge, but at least it makes sense from a set-theoretical point of view.

Now the problem is to find a good distance on the set of abstract compact metric spaces. The natural concept here is the **Gromov–Hausdorff distance**, which is obtained by formally *taking the quotient of the Hausdorff distance by isometries*: If $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ are two compact metric spaces, define

$$d_{GH}(\mathcal{X}, \mathcal{Y}) = \inf\ d_H(\mathcal{X}', \mathcal{Y}'), \tag{27.1}$$

where the infimum is taken over all *isometric embeddings* $\mathcal{X}'$, $\mathcal{Y}'$ of $\mathcal{X}$ and $\mathcal{Y}$ into a common metric space $\mathcal{Z}$; this means that $\mathcal{X}'$ is isometric to $\mathcal{X}$ (more rigorously, belongs to the isometry class defined by $\mathcal{X}$), $\mathcal{Y}'$ is isometric to $\mathcal{Y}$, and both $\mathcal{X}'$ and $\mathcal{Y}'$ are subspaces of $\mathcal{Z}$.

Of course, there is no loss of generality in choosing $\mathcal{Z} = \mathcal{X}' \cup \mathcal{Y}'$, but let me insist: the metric on $\mathcal{X}'$, $\mathcal{Y}'$ has to be the metric induced by $\mathcal{Z}$! In that situation I shall say that $(\mathcal{X}', \mathcal{Y}')$ constitute a **metric coupling** of the abstract spaces $(\mathcal{X}, \mathcal{Y})$. Two metric couplings $(\mathcal{X}', \mathcal{Y}')$ and $(\mathcal{X}'', \mathcal{Y}'')$ will be said to be isometric if there is an isometry $F : (\mathcal{X}' \cup \mathcal{Y}') \to (\mathcal{X}'' \cup \mathcal{Y}'')$ which restricts to isometries $\mathcal{X}' \to \mathcal{X}''$ and $\mathcal{Y}' \to \mathcal{Y}''$.

Representation by semi-distances

As we know, all the probabilistic information contained in a coupling (X, Y) of two probability spaces $(\mathcal{X}, \nu_{\mathcal{X}})$ and $(\mathcal{Y}, \nu_{\mathcal{Y}})$ is summarized by a *joint probability measure on the*

product space $\mathcal{X} \times \mathcal{Y}$. There is an analogous statement for metric couplings: All the geometric information contained in a metric coupling $(\mathcal{X}', \mathcal{Y}')$ of two abstract metric spaces $\mathcal{X}$ and $\mathcal{Y}$ is summarized by a *semi-distance on the disjoint union* $\mathcal{X} \sqcup \mathcal{Y}$. Here are the definitions:

- a semi-distance on a set $\mathcal{Z}$ is a map $d : \mathcal{Z} \times \mathcal{Z} \to [0, +\infty)$ satisfying $d(x, x) = 0$, $d(x, y) = d(y, x)$, $d(x, z) \leq d(x, y) + d(y, z)$, but *not necessarily* $d(x, y) = 0 \Longrightarrow x = y$;

- the disjoint union $\mathcal{X} \sqcup \mathcal{Y}$ is the union of two disjoint isometric copies of $\mathcal{X}$ and $\mathcal{Y}$. The particular way in which this disjoint union is constructed does not matter; for instance, take any representative of $\mathcal{X}$ (still denoted $\mathcal{X}$ for simplicity), any representative of $\mathcal{Y}$, and set $\mathcal{X} \sqcup \mathcal{Y} = (\{0\} \times \mathcal{X}) \cup (\{1\} \times \mathcal{Y})$. Then $\{0\} \times \mathcal{X}$ is isometric to $\mathcal{X}$ via the map $(0, x) \to x$, etc.

Not all semi-distances however are allowed. In a probabilistic context, the only admissible couplings of two measures $\nu_{\mathcal{X}}$ and $\nu_{\mathcal{Y}}$ are those whose joint law π has marginals $\nu_{\mathcal{X}}$ and $\nu_{\mathcal{Y}}$. There is a similar principle for metric couplings: If $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ are two given abstract metric spaces, the only admissible semi-distances on $\mathcal{X} \sqcup \mathcal{Y}$ are those whose restriction to $\mathcal{X} \times \mathcal{X}$ (resp. $\mathcal{Y} \times \mathcal{Y}$) coincides with $d_{\mathcal{X}}$ (resp. $d_{\mathcal{Y}}$). When that condition is satisfied, it will be possible to reconstruct a metric coupling from the semi-distance, by just "taking the quotient" of $\mathcal{X} \sqcup \mathcal{Y}$ by the semi-distance d, in other words deciding that two points a and b with $d(a, b) = 0$ really are the same.

All this is made precise by the following statement:

Proposition 27.1 (Metric couplings as semi-distances). *Let $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ be two disjoint metric spaces, and let $\mathcal{X} \sqcup \mathcal{Y}$ be their union. Then*

(i) Let $(\mathcal{X}', \mathcal{Y}')$ be a metric coupling of $\mathcal{X}$ and $\mathcal{Y}$; let $f : \mathcal{X} \to \mathcal{X}'$ and $g : \mathcal{Y} \to \mathcal{Y}'$ be corresponding isometries, and let $(\mathcal{Z}, d_{\mathcal{Z}})$ be the ambient metric space containing $\mathcal{X}'$ and $\mathcal{Y}'$. Whenever a, b belong in $\mathcal{X} \sqcup \mathcal{Y}$, define

$$d(a, b) = \begin{cases} d_{\mathcal{X}}(a, b) & \text{if } a, b \in \mathcal{X} \\ d_{\mathcal{Y}}(a, b) & \text{if } a, b \in \mathcal{Y} \\ d_{\mathcal{Z}}(f(a), g(b)) & \text{if } a \in \mathcal{X},\ b \in \mathcal{Y} \\ d_{\mathcal{Z}}(g(a), f(b)) & \text{if } a \in \mathcal{Y},\ b \in \mathcal{X}. \end{cases}$$

Then d is a semi-distance on $\mathcal{X} \sqcup \mathcal{Y}$, whose restriction to $\mathcal{X} \times \mathcal{X}$ (resp. $\mathcal{Y} \times \mathcal{Y}$) is $d_{\mathcal{X}}$ (resp. $d_{\mathcal{Y}}$).

(ii) Conversely, let d be a semi-distance on $\mathcal{X} \sqcup \mathcal{Y}$, whose restriction to $\mathcal{X} \times \mathcal{X}$ (resp. $\mathcal{Y} \times \mathcal{Y}$) is $d_{\mathcal{X}}$ (resp. $d_{\mathcal{Y}}$). On $\mathcal{X} \sqcup \mathcal{Y}$, define the relation $\mathcal{R}$ by the property

$$x \,\mathcal{R}\, x' \iff d(x, x') = 0.$$

This is an equivalence relation, so one may define

$$\mathcal{Z} = \mathcal{X} \sqcup \mathcal{Y}/d := \mathcal{X} \sqcup \mathcal{Y}/\mathcal{R}$$

as the set of classes of equivalence of $\mathcal{X} \sqcup \mathcal{Y}$. Write $\overline{x}$ for the equivalence class of x, and define

$$d_{\mathcal{Z}}(\overline{a}, \overline{b}) = d(a, b)$$

(this does not depend on the choice of representatives, but only on the equivalence classes). Then $x \to \overline{x}$ is an isometric embedding of $\mathcal{X}$ into $(\mathcal{Z}, d_{\mathcal{Z}})$, and similarly $y \to \overline{y}$ is an isometric embedding of $\mathcal{Y}$ into $(\mathcal{Z}, d_{\mathcal{Z}})$.

The reader should have no difficulty to write down the proof; just be patient enough and make sure that you consider all the cases!

Now the following property should not come as a surprise:

Theorem 27.2 (Metric Gluing Lemma). *Let $(\mathcal{X}_1, d_1)$, $(\mathcal{X}_2, d_2)$, $(\mathcal{X}_3, d_3)$ be three abstract compact metric spaces. If $(\mathcal{X}_1', \mathcal{X}_2')$ is a metric coupling of $(\mathcal{X}_1, \mathcal{X}_2)$ and $(\mathcal{X}_2'', \mathcal{X}_3'')$ is a metric coupling of $(\mathcal{X}_2, \mathcal{X}_3)$, then there is a triple of metric spaces $(\widetilde{\mathcal{X}}_1, \widetilde{\mathcal{X}}_2, \widetilde{\mathcal{X}}_3)$, all subspaces of a common metric space $(\mathcal{Z}, d_{\mathcal{Z}})$, such that $(\widetilde{\mathcal{X}}_1, \widetilde{\mathcal{X}}_2)$ is isometric (as a coupling) to $(\mathcal{X}_1', \mathcal{X}_2')$, and $(\widetilde{\mathcal{X}}_2, \widetilde{\mathcal{X}}_3)$ is isometric to $(\mathcal{X}_1'', \mathcal{X}_2'')$.*

Sketch of proof. By means of Proposition 27.1, the metric coupling $(\mathcal{X}_1', \mathcal{X}_2')$ may be thought of as a semi-distance d_{12} on $\mathcal{X}_1 \sqcup \mathcal{X}_2$; and similarly, $(\mathcal{X}_2'', \mathcal{X}_3'')$ may be thought of as a semi-distance d_{23} on $\mathcal{X}_2 \sqcup \mathcal{X}_3$. Then, for $x_1 \in \mathcal{X}_1$ and $x_3 \in \mathcal{X}_3$, define

$$d_{13}(x_1, x_3) = \inf_{x_2 \in \mathcal{X}_2} \big[d_{12}(x_1, x_2) + d_{23}(x_2, x_3) \big].$$

Next, on $\mathcal{X}_1 \sqcup \mathcal{X}_2 \sqcup \mathcal{X}_3$ introduce the semi-distance

$$d(a, b) = \begin{cases} d_{12}(a, b) & \text{if } a, b \in \mathcal{X}_1 \sqcup \mathcal{X}_2 \\ d_{23}(a, b) & \text{if } a, b \in \mathcal{X}_2 \sqcup \mathcal{X}_3 \\ d_{13}(a, b) & \text{if } a \in \mathcal{X}_1 \text{ and } b \in \mathcal{X}_3 \\ d_{13}(b, a) & \text{if } a \in \mathcal{X}_3 \text{ and } b \in \mathcal{X}_1. \end{cases}$$

This is a semi-distance; so one can define

$$\mathcal{Z} = (\mathcal{X}_1 \sqcup \mathcal{X}_2 \sqcup \mathcal{X}_3)/d,$$

and repeat the same reasoning as in Proposition 27.1. □

Representation by approximate isometries

If a correspondence $\mathcal{R}$ preserves distances, in the sense that $d(x, x') = d(y, y')$ for all (x, y), (x', y') in $\mathcal{R}$, then it is almost obvious that $\mathcal{R}$ is the graph of an isometry between $\mathcal{X}$ and $\mathcal{Y}$. To measure how far the correspondence is from being an isometry, define its **distortion** by the formula

$$\mathrm{dis}\,(\mathcal{R}) = \sup_{(x,y),(x',y') \in \mathcal{R}} \Big| d_{\mathcal{Y}}(y, y') - d_{\mathcal{X}}(x, x') \Big|.$$

Then it can be shown that

$$d_{GH}(\mathcal{X}, \mathcal{Y}) = \frac{1}{2} \inf \mathrm{dis}\,(\mathcal{R}), \tag{27.2}$$

where the infimum is over all correspondences $\mathcal{R}$ between $\mathcal{X}$ and $\mathcal{Y}$.

There is an even more handy way to consider the Gromov–Hausdorff distance, in terms of **approximate isometries**. By definition, an ε-isometry between $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ is a map $f : \mathcal{X} \to \mathcal{Y}$ that is "almost an isometry", that is:

(a') it almost preserves distances: for all x, x' in $\mathcal{X}$,

$$\Big| d(f(x), f(x')) - d(x, x') \Big| \le \varepsilon;$$

(b') it is almost surjective:

$$\forall y \in \mathcal{Y} \quad \exists x \in \mathcal{X}; \qquad d(f(x), y) \le \varepsilon.$$

In particular, $d_H(f(\mathcal{X}), \mathcal{Y}) \le \varepsilon$.

Remark 27.3. Heuristically, an ε-isometry is a map that you can't distinguish from an isometry if you are short-sighted, that is, if you measure all distances with a possible error of about ε.

It is not clear whether one can reformulate the Gromov–Hausdorff distance in terms of ε-isometries, but at least from the qualitative point of view this works fine: It can be shown that

$$\frac{2}{3}\, d_{GH}(\mathcal{X}, \mathcal{Y}) \;\le\; \inf\Big\{\varepsilon;\ \exists f\ \varepsilon\text{-isometry } \mathcal{X} \to \mathcal{Y}\Big\} \;\le\; 2\, d_{GH}(\mathcal{X}, \mathcal{Y}). \tag{27.3}$$

Indeed, if f is an ε-isometry, define a relation $\mathcal{R}$ by

$$(x, y) \in \mathcal{R} \iff \quad d(f(x), y) \le \varepsilon;$$

then $\mathrm{dis}\,(\mathcal{R}) \le 3\varepsilon$, and the left inequality in (27.3) follows by formula (27.2). Conversely, if $\mathcal{R}$ is a relation with distortion η, then for any $\varepsilon > \eta$ one can define an ε-isometry f whose graph is included in $\mathcal{R}$: The idea is to define $f(x) = y$, where y is such that $(x, y) \in \mathcal{R}$. (See the comments at the end of the bibliographic notes.)

The symmetry between $\mathcal{X}$ and $\mathcal{Y}$ seems to have been lost in (27.3), but this is not serious, because *any approximate isometry admits an approximate inverse*: if f is an ε-isometry $\mathcal{X} \to \mathcal{Y}$, then there is $f' : \mathcal{Y} \to \mathcal{X}$ such that f' is a (4ε)-isometry $\mathcal{Y} \to \mathcal{X}$, and for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$, one has

$$d_{\mathcal{X}}\big(f' \circ f(x),\, x\big) \le 3\varepsilon, \qquad d_{\mathcal{Y}}\big(f \circ f'(y),\, y\big) \le \varepsilon. \tag{27.4}$$

This map f' will be called an *ε-inverse* of f.

To construct f', consider the relation $\mathcal{R}$ induced by f, which has distortion at most 3ε; then flip the components of $\mathcal{R}$ to get a relation $\mathcal{R}'$ from $\mathcal{Y}$ to $\mathcal{X}$, with (obviously) the same distortion as $\mathcal{R}$, and construct a (4ε)-isometry $f' : \mathcal{Y} \to \mathcal{X}$ whose graph is a subset of $\mathcal{R}$. Then $(f(x), x) \in \mathcal{R}'$ and $(f(x), f'(f(x))) \in \mathcal{R}'$, so $d(f'(f(x)), x) \le \mathrm{dis}\,(\mathcal{R}') + d(f(x), f(x)) \le 2\varepsilon$. Similarly, the identity $(f'(y), y) \in \mathcal{R}$ implies $d(f(f'(y)), y) \le \varepsilon$.

If there is an ε-isometry between $\mathcal{X}$ and $\mathcal{Y}$, then I shall say that $\mathcal{X}$ and $\mathcal{Y}$ are ε-isometric. This terminology has the drawback that the order of $\mathcal{X}$ and $\mathcal{Y}$ matters: if $\mathcal{X}$ and $\mathcal{Y}$ are ε-isometric, then $\mathcal{Y}$ and $\mathcal{X}$ are not necessarily ε-isometric; but at least they are (2ε)-isometric, so from the qualitative point of view this is not a problem.

Lemma 27.4 (approximate isometries converge to isometries). *Let $\mathcal{X}$ and $\mathcal{Y}$ be two compact metric spaces, and for each $k \in \mathbb{N}$ let f_k be a ε_k-isometry, where $\varepsilon_k \to 0$. Then, up to extraction of a subsequence, f_k converges to an isometry.*

Sketch of proof of Lemma 27.4. Introduce a dense subset S of $\mathcal{X}$. For each $x \in \mathcal{X}$, the sequence $(f_k(x))$ is valued in the compact set $\mathcal{Y}$, and so, up to extraction of a subsequence, it converges to some $f(x) \in \mathcal{Y}$. By a diagonal extraction, we may assume that $f_k(x) \to f(x)$ for all $x \in \mathcal{X}$. By passing to the limit in the inequality satisfied by f_k, we see that f is distance-preserving. By uniform continuity, it may be extended into a distance-preserving map $\mathcal{X} \to \mathcal{Y}$.

Similarly, there is a distance-preserving map $\mathcal{Y} \to \mathcal{X}$, denoted by g. Then the composition $g \circ f$ is a distance-preserving map $\mathcal{X} \to \mathcal{X}$, and since $\mathcal{X}$ is compact it follows from a well-known theorem that $g \circ f$ is a bijection. As a consequence, both f and g are bijective, so they are isometries. □

Remark 27.5. The above proof establishes the pointwise convergence of (a subsequence of) f_k to f, but in fact one can impose the uniform convergence; see Theorem 27.10.

The Gromov–Hausdorff space

After all these preparations, we may at least understand why d_{GH} is a honest distance:

(i) It is obviously symmetric.

(ii) Let $\mathcal{X}$ and $\mathcal{Y}$ be two metric spaces; define $\mathcal{Z}$ to be the disjoint union $\mathcal{X} \sqcup \mathcal{Y}$, and define a *distance* d on $\mathcal{X} \sqcup \mathcal{Y}$ by letting $d(x, y) = D > 0$ as soon as $(x, y) \in \mathcal{X} \times \mathcal{Y}$. If D is chosen large enough, this is indeed a distance; so the injections $(x, y) \to x$ and $(x, y) \to y$ realize a metric coupling of $(\mathcal{X}, \mathcal{Y})$. It follows that the infimum in (27.1) is not $+\infty$.

(iii) Obviously, $d_{GH}(\mathcal{X}, \mathcal{X}) = 0$. Conversely, if $\mathcal{X}$ and $\mathcal{Y}$ are two abstract compact metric spaces such that $d_{GH}(\mathcal{X}, \mathcal{Y}) = 0$, introduce any two representatives of these isometry classes (still denoted $\mathcal{X}$ and $\mathcal{Y}$ for simplicity), then there is a family of $(1/k)$-isometries $f_k : \mathcal{X} \to \mathcal{Y}$. Up to extraction of a subsequence, this family converges to a true isometry, so $\mathcal{X}$ and $\mathcal{Y}$ are isometric, and define the same isometry class.

(iv) Finally, the triangular inequality follows easily from the metric Gluing Lemma — just as the triangle inequality for the Wasserstein distance was a consequence of the probabilistic Gluing Lemma.

So, if $(\mathcal{GH}, d_{GH})$ stands for the set of all classes of isometry of compact metric spaces, equipped with the Gromov–Hausdorff distance, then this is a complete separable metric space. An explicit countable dense family is provided by the family of all finite subsets of $\mathbb{N}$ with rational-valued distances. Convergence in the Gromov–Hausdorff distance is called **Gromov–Hausdorff convergence**.

It is equivalent to express the Gromov–Hausdorff convergence in terms of embeddings and Hausdorff distance, or in terms of distortions of correspondences, or in terms of approximate isometries. This leads us to the following definition.

Definition 27.6 (Gromov–Hausdorff convergence). *Let $(\mathcal{X}_k)_{k\in\mathbb{N}}$ be a sequence of compact metric spaces, and let $\mathcal{X}$ be a compact metric space. Then it is said that $\mathcal{X}_k$ converges to $\mathcal{X}$ in the Gromov–Hausdorff topology if any one of the three equivalent statements is satisfied:*

(i) $d_{GH}(\mathcal{X}_k, \mathcal{X}) \longrightarrow 0$;

(ii) There are correspondences $\mathcal{R}_k$ between $\mathcal{X}_k$ and $\mathcal{X}$ such that $\operatorname{dis} \mathcal{R}_k \longrightarrow 0$;

(iii) There are ε_k-isometries $f_k : \mathcal{X}_k \to \mathcal{X}$, for some sequence $\varepsilon_k \to 0$.

This convergence will be denoted by $\mathcal{X}_k \underset{GH}{\longrightarrow} \mathcal{X}$, or just $\mathcal{X}_k \longrightarrow \mathcal{X}$.

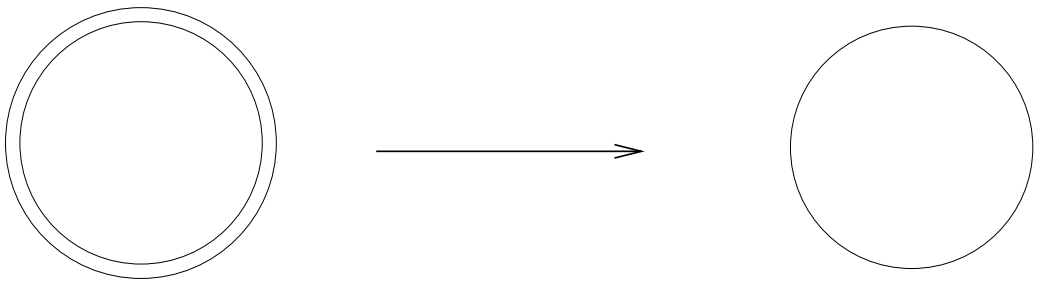


Fig. 27.2. A very thin tyre (2-dimensional manifold) is very close to a circle (1-dimensional manifold)

Remark 27.7. Keeping Remark 27.3 in mind, two spaces are close in Gromov–Hausdorff topology if they look the same to a short-sighted person. I learnt from Lott the expression *Mr. Magoo topology* to convey this idea.

Remark 27.8. It is important to allow the approximate isometries to be discontinuous. Figure 27.3 below shows a simple example where two spaces $\mathcal{X}$ and $\mathcal{Y}$ are very close to each other in Gromov–Hausdorff topology, although there is no continuous map $\mathcal{X} \to \mathcal{Y}$. (Still there is a famous convergence theorem by Gromov showing that such behavior is ruled out by bounds on the curvature.)

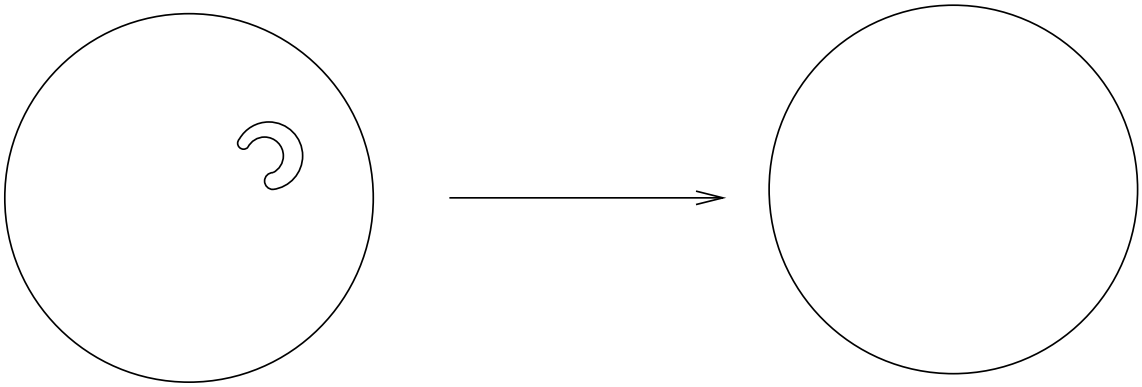


Fig. 27.3. A balloon with a very small handle (not simply connected) is very close to a balloon without handle (simply connected).

The Gromov–Hausdorff topology enjoys the very nice property that *any geometric statements which can be expressed in terms of the distances between a finite number of points automatically passes to the limit.* For example, consider the statement "Any pair (x, y) of points in $\mathcal{X}$ admit a midpoint", which in a complete space is characteristic of a length space. This only involves the distance between configurations of three points (x, y and the candidate midpoint), so it passes to the limit. Then geodesics can be reconstructed from successive midpoints. In this way one can prove statements such as

Theorem 27.9 (convergence of length spaces). *Let $(\mathcal{X}_k)$ be a sequence of compact length spaces converging to $\mathcal{X}$ in Gromov–Hausdorff topology; then $\mathcal{X}$ is a length space. Moreover, if f_k is an ε_k-isometry $\mathcal{X}_k \to \mathcal{X}$, and γ_k is a geodesic curve in $\mathcal{X}_k$ such that $f_k \circ \gamma_k$ converges to some curve γ in $\mathcal{X}$, then γ is a geodesic.*

Gromov–Hausdorff topology and nets

Given $\varepsilon > 0$, a set $\mathcal{N}$ in a metric space $(\mathcal{X}, d)$ is called an ε-net (in $\mathcal{X}$) if the enlargement $S^{\varepsilon]}$ covers $\mathcal{X}$; in other words, for any $x \in \mathcal{X}$ there is $y \in \mathcal{N}$ such that $d(x, y) \le \varepsilon$.

If $\mathcal{N}$ is an ε-net in $\mathcal{X}$, clearly the distance between $\mathcal{N}$ and $\mathcal{X}$ is at most ε. And if $\mathcal{X}$ is compact, then it admits finite ε-nets for all $\varepsilon > 0$, so it can be approximated in Gromov–Hausdorff topology by a sequence of finite sets.

In fact, it is another nice feature of the Gromov–Hausdorff topology that it ultimately *always* reduces to convergence of finite ε-nets. More precisely, $\mathcal{X}_n \longrightarrow \mathcal{X}$ in the Gromov–Hausdorff topology if and only if for any $\varepsilon > 0$ there exists a finite ε-net $\{x_1, \dots, x_k\}$

in $\mathcal{X}$, and for n large enough there is an ε-net $\{x_1^{(n)}, \dots, x_k^{(n)}\}$ in $\mathcal{X}_n$, and for all $j \le k$, $x_j^{(n)} \longrightarrow x_j$.

This leads to the main criterion of compactness in Gromov–Hausdorff topology. Recall that a metric space $\mathcal{X}$ is said to be **totally bounded** if for any $\varepsilon > 0$ it can be covered by a finite number of balls of radius ε. If $N(\varepsilon)$ is the minimal number of such balls, then $\varepsilon \longmapsto N(\varepsilon)$ can be found as a "modulus of total boundedness". Then the following statement, due to Gromov, is vaguely reminiscent of Ascoli's theorem:

Theorem 27.10 (compactness criterion in Gromov–Hausdorff topology). *A family $\mathcal{F}$ of compact metric spaces is precompact in the Gromov–Hausdorff topology if and only if it is uniformly totally bounded, in the sense that for any $\varepsilon > 0$ there is $N = N(\varepsilon)$ such that for any $\mathcal{X} \in \mathcal{F}$ there is an ε-net of cardinality at most N.*

Noncompact spaces

There is no problem in extending the definition of the Gromov–Hausdorff distance to noncompact spaces, except of course that the resulting "distance" might be infinite. But even when it is finite, this notion is of limited use. A good analogy is the concept of uniform convergence of functions, which usually is too strong a notion for noncompact spaces, and should be replaced by *locally uniform* convergence, i.e. uniform convergence on each compact subset.

At first sight, it does not seem to make much sense to define a notion of local Gromov–Hausdorff convergence. Indeed, if a sequence $(\mathcal{X}_k)_{k\in\mathbb{N}}$ of metric spaces is given, there is a priori no canonical choice of family of compact sets in $\mathcal{X}_k$ that we can use to compare to a family of compact sets in $\mathcal{X}$. So we had better impose the existence of these compact sets on *each member of the family*. The idea is to exhaust the space $\mathcal{X}$ by compact sets $K^{(\ell)}$ in such a way that each $K^{(\ell)}$ (equipped with the metric induced by $\mathcal{X}$) is a Gromov–Hausdorff limit of corresponding compact sets $K_k^{(\ell)} \subset \mathcal{X}_k$ (each of them with the induced metric), as $k \to \infty$.

Definition 27.11 (local Gromov–Hausdorff convergence). *Let $(\mathcal{X}_k)_{k\in\mathbb{N}}$ be a family of Polish spaces, and let $\mathcal{X}$ be another Polish space. It is said that $\mathcal{X}_k$ converges to $\mathcal{X}$ in the local Gromov–Hausdorff topology if there are increasing sequences of compact sets $(K_k^{(\ell)})_{\ell\in\mathbb{N}}$ in $\mathcal{X}_k$, and $(K^{(\ell)})$ in $\mathcal{X}$, such that*

(i) $\bigcup_\ell K^{(\ell)}$ is dense in $\mathcal{X}$;

(ii) for each fixed ℓ, $K_k^{(\ell)}$ converges in Gromov–Hausdorff sense to $K^{(\ell)}$, as $k \to \infty$.

If one works in length spaces, as will be the case in the rest of these notes, the above definition does not seem so good because $K^{(\ell)}$ will in general not be a strictly intrinsic length space: geodesics joining elements of $K^{(\ell)}$ might very well leave $K^{(\ell)}$ at some intermediate time; so properties involving geodesics might not pass to the limit. This is the reason for requirement (iii) in the following definition.

Definition 27.12 (geodesic local Gromov–Hausdorff convergence). *Let $(\mathcal{X}_k, d_k)_{k\in\mathbb{N}}$ be a family of Polish strictly intrinsic length spaces, and let $(\mathcal{X}, d)$ be a Polish space. It is said that $\mathcal{X}_k$ converges to $\mathcal{X}$ in the geodesic local Gromov–Hausdorff topology if there are increasing sequences of compact sets $(K_k^{(\ell)})_{\ell\in\mathbb{N}}$ in $\mathcal{X}_k$, and $(K^{(\ell)})_{\ell\in\mathbb{N}}$ in $\mathcal{X}$, such that (i) and (ii) in Definition 27.11 are satisfied, and in addition*

(iii) For each ℓ, there is ℓ' such that all geodesics joining points of $K_k^{(\ell)}$ have their image contained in $K_k^{(\ell')}$.

Then $(\mathcal{X}, d)$ is automatically a strictly intrinsic length space.

If $\mathcal{X}$ is a boundedly compact space (that is, all closed balls are compact), then there is a kind of natural choice for an exhaustive family of compact sets in $\mathcal{X}$: Pick up an arbitrary point $\star \in \mathcal{X}$, and consider the closed balls $B(\star, R_\ell])$, where R_ℓ is any sequence of positive real numbers going to infinity, say $R_\ell = \ell$. One can fix the sequence R_ℓ once for all, and then the notion of convergence only depends on the choice of the "reference point" or "base point" $\star$ (the point from which the convergence is seen). This suggests that the basic objects should not be just metric spaces, but rather **pointed metric spaces**. By definition, a pointed metric space consists of a triple $(\mathcal{X}, d, \star)$, where $(\mathcal{X}, d)$ is a metric space and $\star$ is some point in $\mathcal{X}$. Sometimes I shall just write $(\mathcal{X}, \star)$ or even just $\mathcal{X}$ as a shorthand for the triple $(\mathcal{X}, d, \star)$.

It is equivalent for a length space to be boundedly compact or to be locally compact; so in the sequel the basic regularity assumption, when considering pointed Gromov–Hausdorff convergence, will be **local compactness**.

All the notions that were introduced in the previous section can be generalized in a completely obvious way to pointed metric spaces: A pointed isometry between $(\mathcal{X}, \star_{\mathcal{X}})$ and $(\mathcal{Y}, \star_{\mathcal{Y}})$ is an isometry which sends $\star_{\mathcal{X}}$ to $\star_{\mathcal{Y}}$; the pointed Gromov–Hausdorff distance d_{pGH} between two pointed spaces $(\mathcal{X}, \star_{\mathcal{X}})$ and $(\mathcal{Y}, \star_{\mathcal{Y}})$ is obtained as an infimum of Hausdorff distances over all pointed isometric embeddings; a pointed correspondence is a correspondence such that $\star_{\mathcal{X}}$ is in correspondence with $\star_{\mathcal{Y}}$; a pointed ε-isometry is an ε-isometry which sends $\star_{\mathcal{X}}$ to $\star_{\mathcal{Y}}$, etc. Then Definition 27.6 can be trivially transformed into a pointed notion of convergence, expressing the fact that *for each R, the closed ball $B(\star_k, R])$ in $\mathcal{X}_k$ converges to the closed ball $B(\star, R])$ in $\mathcal{X}$*. By an easy extraction argument, this is equivalent to the following alternative definition.

Definition 27.13 (pointed Gromov–Hausdorff convergence). *Let $(\mathcal{X}_k, \star_k)$ be a sequence of pointed locally compact Polish length spaces, and let $(\mathcal{X}, \star)$ be a pointed locally compact Polish space. Then it is said that $\mathcal{X}_k$ converges to $\mathcal{X}$ in the pointed Gromov–Hausdorff topology if any one of the three equivalent statements is satisfied:*

(i) There is a sequence $R_k \to \infty$ such that $d_{pGH}\Big(B(\star_k, R_k]),\, B(\star, R_k])\Big) \longrightarrow 0$;

(ii) There is a sequence $R_k \to \infty$, and there are pointed correspondences $\mathcal{R}_k$ between $B(\star_k, R_k])$ and $B(\star, R_k])$ such that $\mathrm{dis}\,(\mathcal{R}_k) \longrightarrow 0$;

(iii) There are sequences $R_k \to \infty$ and $\varepsilon_k \to 0$, and pointed ε_k-isometries $f_k : B(\star_k, R_k]) \to B(\star, R_k])$ with $\varepsilon_k \to 0$.

Remark 27.14. This notion of convergence implies the geodesic local convergence, as defined in Definition 27.12. Indeed, (i) and (ii) of Definition 27.11 are obviously satisfied, and (iii) follows from the fact that if a geodesic curve has its endpoints in $B_{R]}(\star)$, then its image lies entirely inside $B_{R']}(\star)$ with $R' = 2R$.

Example 27.15 (blow-up). Let M be a Riemannian manifold of dimension n, and x a point in M. For each k, consider the pointed metric space $\mathcal{X}_k = (M, kd, x)$, where x is the basepoint, and kd is just the original geodesic distance on M, dilated by a factor k. Then $\mathcal{X}_k$ converges in the pointed Gromov–Hausdorff topology to the tangent space $T_x M$, pointed at x and equipped with the metric g_x (it is a Euclidean space). This is true as soon as M is just differentiable at x, in the sense of the existence of the tangent space.

Example 27.16. More generally, if $\mathcal{X}$ is a given metric space, and x is a point in $\mathcal{X}$, one can define the rescaled pointed spaces $\mathcal{X}_k = (\mathcal{X}, kd, x)$; if this sequence converges in the pointed Gromov–Hausdorff topology to some metric space $\mathcal{Y}$, then $\mathcal{Y}$ is said to be the tangent space, or **tangent cone**, to $\mathcal{X}$ at x. In many cases, this space coincides with the metric cone built on some length space Σ, which itself can be thought of as the space of tangent directions. (By definition, if (B, d) is a length space, the metric cone over B is obtained by considering $B \times [0, \infty)$, gluing together all the points in the fiber $B \times \{0\}$, and equipping the resulting space with the *cone metric*: $d_c((x,t),(y,s)) = \sqrt{t^2 + s^2 - 2ts \cos d(x,y)}$ when $d(x,y) \leq \pi$, $d_c((x,t),(y,s)) = t + s$ when $d(x,y) > \pi$.)

Example 27.17. For any $p \in [1, \infty)$, define the ℓ^p norm on $\mathbb{R}^n$ by the usual formula $\|x\|_{\ell^p} = (\sum |x_i|^p)^{1/p}$; and let $\mathcal{X}_p$ be the space $\mathbb{R}^n$, equipped with the ℓ^p norm, pointed at 0. Then, as $p \to \infty$, $\mathcal{X}_p$ converges in the pointed Gromov–Hausdorff to $\mathcal{X}_\infty$, which is $\mathbb{R}^n$ equipped with the ℓ^∞ norm, $\|x\|_{\ell^\infty} = \sup |x_i|$. In $\mathcal{X}_p$, geodesics are segments of the form $(1-t)\, a + t\, b$, in particular they are nonbranching (two distinct geodesics cannot coincide on a nontrivial time interval), and unique (any two points are joined by a unique geodesic path). In contrast, geodesics in $\mathcal{X}_\infty$ are branching and definitely nonunique (any two distinct points can be joined by an uncountable set of geodesic paths). We see on this example that neither the nonbranching property, nor the property of uniqueness of geodesics, are preserved under Gromov–Hausdorff convergence. In particular, the huge majority of geodesics $\mathcal{X}_\infty$ cannot be realized as limits of geodesics on $\mathcal{X}_p$.

Remark 27.18. Consider pointed length spaces $(\mathcal{X}_k, \star_k)$ and $(\mathcal{X}, \star)$, and let f_k be a pointed ε_k-isometry $B_{R_k]}(\star_k) \to B_{R_k]}(\star)$. Let then $R'_k \leq R_k$. It is clear that the distortion of f_k on $B_{R'_k]}(\star_k)$ is no more than the distortion of f_k on $B_{R_k]}(\star_k)$. Also if x belongs to $B_{R'_k]}(\star)$, and $\mathcal{X}$ is a length space, then there is $x' \in B_{R'_k]}(\star)$ with $d(x,x') = 2\varepsilon_k$ and $d(\star, x') \leq R'_k - 2\varepsilon_k$; then there is $x'_k \in B_{R_k}(\star_k)$ such that $d(f_k(x'_k), x_k) \leq \varepsilon_k$, so $d(\star, f_k(x'_k)) \leq R'_k - \varepsilon_k$, and then by the distortion property $d(\star_k, x'_k) \leq R'_k - \varepsilon_k + \varepsilon_k = R'_k$; on the other hand, $d(x'_k, x) \leq 2\varepsilon_k + \varepsilon_k = 3\varepsilon_k$. The conclusion is that the restriction of f_k to $B_{R'_k}(\star_k)$ defines a $(3\varepsilon_k)$-isometry $B_{R'_k}(\star_k) \to B_{R'_k}(\star)$. In other words, it is always possible to *reduce* R_k and keeping approximate isometries, provided that one changes ε_k for $3\varepsilon_k$.

Remark 27.19 (important). In the theory of Gromov–Hausdorff convergence, it is often imposed that $R_k = (\varepsilon_k)^{-1}$ in Definition 27.13. This is consistent with Example 27.15, and also with most tangent cones that are usually encountered. I shall however not do so in these notes.

Functional analysis on Gromov–Hausdorff converging sequences

Many theorems that hold true for any metric space, still hold true, after appropriate modification, for any *family* of metric spaces, provided that this family converges in Gromov–Hausdorff sense. Such is the case for some of the basic compactness theorems in functional analysis: Ascoli's theorem and Prokhorov's theorem. I shall not need these results outside the setting of compact spaces, so I shall be sketchy about their formulation in the noncompact case; the reader can easily fill in the gaps.

Proposition 27.20 (Ascoli's theorem on a Gromov–Hausdorff converging sequence). *Let $(\mathcal{X}_k)_{k\in\mathbb{N}}$ be a sequence of compact metric spaces, converging in the Gromov–Hausdorff topology to some compact metric space $\mathcal{X}$, by means of ε_k-approximations $f_k : \mathcal{X}_k \to \mathcal{X}$, admitting approximate inverses f'_k; and let $(\mathcal{Y}_k)_{k\in\mathbb{N}}$ be another sequence*

of compact metric spaces converging to $\mathcal{Y}$ in the Gromov–Hausdorff topology, by means of ε_k-approximations $g_k : \mathcal{Y}_k \to \mathcal{Y}$. Let $(\alpha_k)_{k\in\mathbb{N}}$ be a sequence of maps $\mathcal{X}_k \to \mathcal{Y}_k$ that are asymptotically equicontinuous, in the sense that for every $\varepsilon > 0$, there are $\delta = \delta(\varepsilon) > 0$ and $N = N(\varepsilon) \in \mathbb{N}$ so that for all $k \geq N$,

$$d_{\mathcal{X}_k}(x_k, x'_k) \leq \delta \qquad \Longrightarrow \qquad d_{\mathcal{Y}_k}\big(\alpha_k(x_k), \alpha_k(x'_k)\big) \leq \varepsilon. \tag{27.5}$$

Then after passing to a subsequence, the maps $g_k \circ \alpha_k \circ f'_k : \mathcal{X} \to \mathcal{Y}$ converge uniformly to a continuous map $\alpha : \mathcal{X} \to \mathcal{Y}$.

This statement extends to locally compact spaces converging in the pointed Gromov–Hausdorff topology, and locally asymptotically uniformly equicontinuous maps, provided that the conclusion is weakened into locally uniform convergence.

Remark 27.21. In the conclusion of Proposition 27.20, the maps $g_k \circ \alpha_k \circ f'_k$ may be discontinuous, yet they will converge uniformly.

Proposition 27.22 (Prokhorov's theorem on a Gromov–Hausdorff converging sequence). *Let $(\mathcal{X}_k)_{k\in\mathbb{N}}$ be a sequence of compact metric spaces, converging in the Gromov–Hausdorff topology to some compact metric space $\mathcal{X}$, by means of ε_k-approximations $f_k : \mathcal{X}_k \to \mathcal{X}$. For each k, let μ_k be a probability measure on $\mathcal{X}_k$. Then, after extraction of a subsequence, the sequence $((f_k)_\# \mu_k)_{k\in\mathbb{N}}$ converges weakly to a probability measure μ on $\mathcal{X}$.*

This statement extends to Polish spaces converging by means of local Gromov–Hausdorff approximations, provided that the probability measures μ_k are uniformly tight with respect to the sequences $(K_k^{(\ell)})$ appearing in the definition of local Gromov–Hausdorff approximation.

Here is another simple compactness criterion for which I shall provide a proof.

Proposition 27.23 (compactness of locally bounded measures). *Let $(\mathcal{X}_k, d_k, \star_k)_{k\in\mathbb{N}}$ be a sequence of pointed locally compact Polish spaces converging in the pointed Gromov–Hausdorff topology to some pointed locally compact Polish space $(\mathcal{X}, d, \star)$, by means of pointed ε_k-isometries f_k with $\varepsilon_k \to 0$. Assume that for each R, there is a finite constant $M = M(R)$ such that*

$$\forall k \in \mathbb{N}, \qquad \nu_k[B_{R]}(\star_k)] \leq M.$$

Then, there is a locally finite measure ν such that, up to extraction of a subsequence,

$$(f_k)_\# \nu_k \xrightarrow[k\to\infty]{} \nu$$

in the weak-$*$ convergence of measures (that is, convergence against compactly supported continuous functions).

Proof. Let $R > 0$; then $(f_k)_\# \nu_k[B_{R]}(\star)] = \nu_k[(f_k)^{-1}(B_{R]}(\star)] \leq \nu_k[B_{R+\varepsilon_k}(\star_k)]$ is uniformly bounded by $M(R+1)$ for k large enough. Since on the other hand $B_{R]}(\star)$ is compact, we may extract a subsequence in k such that $(f_k)_\# \nu_k[B_{R]}(\star)]$ converges to some finite measure ν_R in the weak-$*$ topology of $B_{R]}(\star)$. Then the result follows by taking $R = \ell \to \infty$ and applying a diagonal extraction. □

Remark 27.24. There is an easy extension of Proposition 27.23 to geodesic local measured Gromov–Hausdorff convergence.

Adding the measure

Now let us switch from *metric spaces* to *metric-measure spaces*, that are triples $(\mathcal{X}, d, \nu)$, where d is a distance on $\mathcal{X}$ and ν a Borel measure on $\mathcal{X}$. (For brevity I shall sometimes write just $\mathcal{X}$ instead of $(\mathcal{X}, d, \nu)$.) So the question is to measure how far two metric-measure spaces $\mathcal{X}$ and $\mathcal{Y}$ are from each other.

There is a nontrivial choice to be made:

(a) Either we insist that metric-measure spaces are metric spaces in the first place, so that two metric-measure spaces should be declared close only if they are close both in terms of the metric and in terms of the measure;

(b) Or we think that only the measure is relevant, and we should disregard sets of zero or small measure when estimating how far two metric-measure spaces are.

In the first case, one should identify two spaces $(\mathcal{X}, d, \nu)$ and $(\mathcal{X}', d', \nu')$ only if they are *isomorphic* as metric-measure spaces, which means that there exists a measurable bijection $f : \mathcal{X} \to \mathcal{X}'$ such that f is an isometry, and f preserves the measure: $f_\# \nu = \nu'$. Such a map is naturally called a **measure-preserving isometry**, and its inverse f^{-1} is automatically a measure-preserving isometry. (Note: It is not enough to impose that $\mathcal{X}$ and $\mathcal{X}'$ are isometric as metric spaces, and isometric as measure spaces: the same map should do the job for both isomorphisms.)

In the second case, one should identify sets that are isomorphic up to a zero measure set; so a natural thing to do is to declare that $(\mathcal{X}, d, \nu)$ and $(\mathcal{X}', d', \nu')$ are the same if there is a measure-preserving isometry between $\mathrm{Spt}\, \nu$ and $\mathrm{Spt}\, \nu'$, seen as subspaces of $\mathcal{X}$ and $\mathcal{X}'$ respectively.

Here below is illustrated a classical example of a convergence which holds true in the sense of (b), not in the sense of (a):

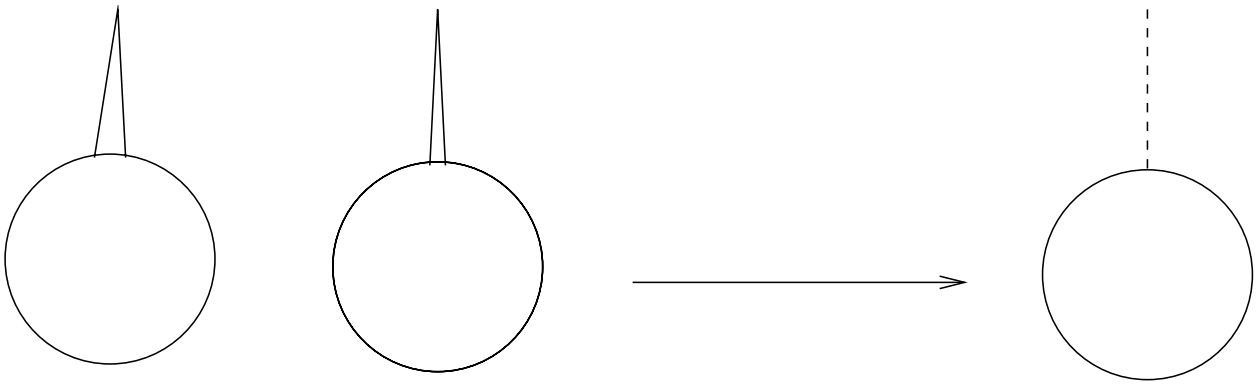


Fig. 27.4. Here is an example of "reduction of support" that can arise in measured Gromov–Hausdorff convergence. This is a balloon with a very thin spike; in Gromov–Hausdorff sense it is approximated by a balloon to which a one-dimensional spike is attached, that carries no measure.

Now it is easy to cook up notions of distance between metric-measure spaces. For a start, let us restrict to compact probability spaces. Pick up a distance which metrizes the weak topology on the set of probability measures, such as the Prokhorov distance d_P, and introduce the **Gromov–Hausdorff–Prokhorov distance** by the formula

$$d_{GHP}(\mathcal{X}, \mathcal{Y}) = \inf \Big\{ d_H(\mathcal{X}', \mathcal{Y}') + d_P(\nu_{\mathcal{X}'}, \nu_{\mathcal{Y}'}) \Big\},$$

where the infimum is taken over all measure-preserving isometric embeddings $f : (\mathcal{X}, \nu_{\mathcal{X}}) \to (X', \nu_{\mathcal{X}'})$ and $g : (\mathcal{Y}, \nu_{\mathcal{Y}}) \to (\mathcal{Y}', \nu_{\mathcal{Y}'})$ into a *common* metric space $\mathcal{Z}$. That choice would

correspond to point of view (a), while in point of view (b) one would rather use the **Gromov–Prokhorov distance**, which is defined, with the same notation, as

$$d_{GP}(\mathcal{X}, \mathcal{Y}) = \inf\ d_P(\nu_{\mathcal{X}'}, \nu_{\mathcal{Y}'}).$$

Both d_{GHP} and d_{GP} satisfy the triangular inequality, as can be checked by a gluing argument again. (Now one should use both the metric and the probabilistic gluing!) Then there is no difficulty in checking that d_{GHP} is a honest distance on classes of metric-measure isomorphisms, with point of view (a). Similarly, d_{GP} is a distance on classes of metric-measure isomorphisms, with point of view (b), but now it is quite nontrivial to check that $[d_{GP}(\mathcal{X}, \mathcal{Y}) = 0] \Longrightarrow [\mathcal{X} = \mathcal{Y}]$. I shall not insist on this point, for in the sequel I shall focus on point of view (a).

There are several variants of these constructions:

1. *Use other distances on probability metrics.* Essentially everybody agrees on the Hausdorff distance to measure distances between sets, but as we know, there are many natural choices of distances between probability measures. In particular, one can replace the Prokhorov distance by the Wasserstein distance of order p, and thus obtain the **Gromov–Hausdorff–Wasserstein distance of order** p:

$$d_{GHW_p}(\mathcal{X}, \mathcal{Y}) = \inf \Big\{ d_H(\mathcal{X}', \mathcal{Y}') + W_p(\nu_{\mathcal{X}'}, \nu_{\mathcal{Y}'}) \Big\},$$

and of course the **Gromov–Wasserstein distance of order** p:

$$d_{GW_p}(\mathcal{X}, \mathcal{Y}) = \inf\ W_p(\nu_{\mathcal{X}'}, \nu_{\mathcal{Y}'}).$$

2. *Measure distances between spaces on which the measure is finite but not necessarily normalized to 1.* This obviously amounts to measure distances between finite nonnegative measures that are not necessarily normalized. There are two rather natural strategies (and many variants). The first one consists in using the bounded Lipschitz distance, as defined in (6.6), which makes sense for arbitrary signed measures; in this way one can define the "Gromov–Hausdorff–bounded-Lipschitz distance" d_{GHbL} and the "Gromov–bounded-Lipschitz distance" d_{GbL}, the definitions of which should be obvious to the reader. Another possibility consists in comparing the normalized metric spaces, and then add a penalty that takes into account the discrepancy between the total masses. For instance, if μ and ν are defined on a common space $\mathcal{Z}$, one may let

$$d(\mu, \nu) = d_P\left(\frac{\mu}{\mu[\mathcal{Z}]}, \frac{\nu}{\nu[\mathcal{Z}]}\right) \ + \ \Big|\mu[\mathcal{Z}] - \nu[\mathcal{Z}]\Big|.$$

One may also replace d_P by W_p, or whatever; and change the penalty (why not something like $|\log(\mu[\mathcal{Z}]/\nu[\mathcal{Z}])|$?); etc. So there is a tremendous number of "natural" possibilities.

3. *Consider noncompact spaces.* Here also, there are many possible frameworks, and the reader is free to consider this variety as a wealth or as a big mess. A first possibility is to just ignore the fact that spaces are noncompact: this is not reasonable if one sticks to philosophy (a), because the Hausdorff distance between noncompact spaces is too rigid; but it makes perfect sense with philosophy (b), at least for *finite measures*. Then, one may apply distances d_{GHW_p} to noncompact situations, or variants designed to handle measures that are not probability measures. But when the measures are only σ-finite, this simple approach has to be modified. A possibility consists in localizing as in Definition 27.11.

Another possibility, which makes sense in a locally compact context, consists in pointing as in Definition 27.13 (and one may also impose the same condition as in Remark 27.19).

About the regularity of the measures: In the sequel, when $(\mathcal{X}, d, \nu)$ is a length space equipped with a measure, I shall always implicitly assume that ν is nonzero, and that it is

- σ-finite if $(\mathcal{X}, d)$ is only assumed to be Polish;
- locally finite if $(\mathcal{X}, d)$ is assumed in addition to be locally compact.

(Given a length space $(\mathcal{X}, d)$ and an arbitrary point $\star \in \mathcal{X}$, $\mathcal{X}$ is the union of the closed balls $B_{k]}(\star)$, so a locally finite measure on $\mathcal{X}$ is automatically σ-finite.)

Doubling

The discussion of the previous section showed that one should be cautious about which notion of convergence is used. However, whenever they are available, *doubling estimates*, in the sense of Definition 18.1, basically rule out the discrepancy between approaches (a) and (b) above. The idea is that doubling prevents the formation of sharp spikes as in Figure 27.4. This discussion is not so clearly made in the literature that I know, so in this section I shall provide more careful proofs.

Proposition 27.25 (the metric and metric-measure approach coincide in presence of doubling). *Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two compact Polish probability spaces with diameter at most R. Assume that both $\mathcal{X}$ and $\mathcal{Y}$ are doubling with a constant D. Then*

$$d_{GP}(\mathcal{X}, \mathcal{Y}) \le d_{GHP}(\mathcal{X}, \mathcal{Y}) \le \Phi_{R,D}\big(d_{GP}(\mathcal{X}, \mathcal{Y})\big), \tag{27.6}$$

where

$$\Phi_{R,D}(\delta) = \max\Big(8\,\delta,\, R\,(16\,\delta)^{\frac{1}{\log_2 D}}\Big) + \delta$$

is a function that goes to 0 as $\delta \to 0$, at a rate that is controlled in terms of just upper bounds on R and D.

Proof of Proposition 27.25. The inequality on the left of (27.6) is trivial, so we focus on the right inequality. To start with, let $x \in \mathcal{X}$, $\varepsilon > 0$, then

$$1 = \mu[\mathcal{X}] = \mu[B_{R]}(x)] \le D^N\, \mu[B_{\varepsilon/4]}(x)],$$

where

$$N = \left\lceil \log_2 \frac{4R}{\varepsilon} \right\rceil \le \log_2 \frac{R}{\varepsilon} + 3,$$

and after a few manipulations this leads to

$$\mu[B_{\varepsilon/4]}(x)] \ge \frac{1}{8} \left(\frac{\varepsilon}{R}\right)^{\log_2 D}.$$

Now let $(\mathcal{X}', \mu')$ and $(\mathcal{Y}', \nu')$ be two isomorphic copies of $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ in some metric space $\mathcal{Z}$. Let ε be the Hausdorff distance between $\mathcal{X}'$ and $\mathcal{Y}'$, and δ the Prokhorov distance between μ' and ν'; the goal is to control $\varepsilon + \delta$ in terms of δ alone. If $\varepsilon \le 8\delta$, then we are done, so we might assume that $\delta < \varepsilon/8$. Since $\varepsilon > 0$, there is, say, some $x \in \mathcal{X}'$ such that the ball $B_{\varepsilon/2]}(x)$ does not intersect $\mathcal{Y}'$.

Now, the doubling property of $(\mathcal{X}, \mu)$ is of course transferred to $(\mathcal{X}', \mu')$, so by the previous estimate

$$\mu'[B_{\varepsilon/4]}(x)] \ge \frac{1}{8} \left(\frac{\varepsilon}{R}\right)^{\log_2 D}. \tag{27.7}$$

By definition of the Prokhorov distance,

$$\mu'[B_{\varepsilon/4]}(x)] \le \nu'[B_{\varepsilon/4+2\delta}(x)] + 2\delta. \tag{27.8}$$

From (27.7) and (27.8) it follows that

$$\nu'[B_{\varepsilon/4+2\delta]}(x)] \ge \frac{1}{8}\left(\frac{\varepsilon}{R}\right)^{\log_2 D} - 2\delta. \tag{27.9}$$

Since $\delta < \varepsilon/8$, the ball $B_{\varepsilon/4+2\delta]}$ is included in the ball $B_{\varepsilon/2]}(x)$, which does not intersect $\mathcal{Y}'$; so the left-hand side in (27.9) has to be 0. It follows that

$$\varepsilon \le R\,(16\delta)^{\frac{1}{\log_2 D}};$$

and then the conclusion easily follows. □

The conclusion is better appreciated in view of the result of the following exercise.

Exercise 27.26. Let $(\mathcal{X}_k, d_k, \nu_k)$ be a sequence of Polish probability spaces converging in the sense of d_{GP} to $(\mathcal{X}, d, \nu)$. Assume that each ν_k is doubling, with a uniform bound on the doubling constant. Prove that ν is also doubling.

The combination of Proposition 27.25 and Exercise 27.26 yields the following corollary:

Corollary 27.27 (d_{GP} convergence and doubling imply d_{GHP} convergence). *Let $(\mathcal{X}_k, d_k, \nu_k)$ be a family of Polish probability spaces satisfying a uniform doubling condition, uniformly bounded, and converging to $(\mathcal{X}, d, \nu)$ in Gromov–Prokhorov sense. Then $(\mathcal{X}_k, d_k, \nu_k)$ also converges in Gromov–Hausdorff sense to $(\mathcal{X}, d, \nu)$. In particular, $(\mathcal{X}_k, d_k)$ converges to $(\mathcal{X}, d)$ in Gromov–Hausdorff sense.*

This corollary basically says that from the qualitative point of view, the distinction between points of view (a) and (b) is nonessential when dealing with the convergence of probability spaces satisfying a uniform doubling estimate. A more careful discussion would extend this conclusion to metric-measure spaces that are not necessarily probability spaces; and then to the pointed convergence of metric-measure spaces, provided that the doubling constant on a ball of radius R (around the base point of each space) only depends on R.

When doubling estimates are not available, things are not so simple and it does matter whether one adheres to philosophy (a) or (b). Point of view (b) is the one that was mainly developed by Gromov, in relation with the phenomenon of *concentration of measure*. It is also the point of view that was adopted by Sturm in his study of the stability of displacement convexity. Nevertheless, I shall prefer to stick here to point of view (a), partly because this is the approach which Lott and myself adopted for the study of the stability of optimal transport, partly because it can be argued that point of view (a) provides a more precise notion of convergence and description of the limit space. For instance, in the example of Figure 27.4, the fact that the limit space has a spike carrying zero measure retains information about the asymptotic shape of the sequence of spaces. This will of course not prevent me from throwing away pieces with zero measure by restricting to the support of the measure, when that is possible.

Doubling has another use in the present context: It leads to uniform total boundedness estimates, and therefore to compactness statements via Theorem 27.10.

Proposition 27.28. *Let $(\mathcal{X}, d)$ be a metric space with diameter bounded above by R, equipped with a finite (nonzero) D-doubling measure ν. Then for any $\varepsilon > 0$ there is a number $N = N(\varepsilon)$, only depending on R and D, such that $\mathcal{X}$ can be covered with N balls of radius ε.*

Proof of Proposition 27.28. Without loss of generality, we might assume that $\nu[\mathcal{X}] = 1$. Let $r = \varepsilon/2$, and let n be such that $R \leq 2^n r$. Choose an arbitrary point x_1 in $\mathcal{X}$; then a point x_2 in $\mathcal{X} \setminus (B_{2r}(x_1))$, a point x_3 in $\mathcal{X} \setminus (B_{2r}(x_1) \cup B_{2r}(x_2))$, and so forth. Then all the balls $B_{r(x_j)}$ are disjoint, and by the doubling property each of them has measure at least D^{-n}. So $\mathcal{X} \setminus (B_r(x_0) \cup \ldots \cup B_r(x_k))$ has measure at most $1 - kD^{-n}$, and therefore D^n is an upper bound on the number of points x_j that can be chosen.

Now let $x \in \mathcal{X}$. There is at least one index j such that $d(x, x_j) < 2r$; otherwise x would lie in the complement of the union of all the balls $B_{2r}(x_j)$, and could be added to the family $\{x_j\}$. So $\{x_j\}$ constitutes a $2r$-net in $\mathcal{X}$, with cardinality at most $N = D^n$. This concludes the proof. □

Measured Gromov–Hausdorff topology

After all this discussion I can state a precise definition of the notion of convergence that will be used in the sequel for metric-measure spaces: this is the **measured Gromov–Hausdorff topology**. It is associated with the convergence of spaces as metric spaces *and* as measure spaces. This concept can be defined quantitatively in terms of, e.g., the distance d_{GHP} and its variants, but I shall be content with a purely topological (qualitative) definition. As in the case of the plain Gromov–Hausdorff topology, there is a convenient reformulation in terms of approximate isometries.

Definition 27.29 (measured Gromov–Hausdorff topology). *Let $(\mathcal{X}_k, d_k, \nu_k)_{k\in\mathbb{N}}$, and $(\mathcal{X}, d, \nu)$ be compact metric spaces, equipped with finite nonzero measures. It is said that $\mathcal{X}_k$ converges to $\mathcal{X}$ in the measured Gromov–Hausdorff topology if there are measurable ε_k-isometries $f_k : \mathcal{X}_k \to \mathcal{X}$ such that $\varepsilon_k \to 0$ and*

$$(f_k)_\# \nu_k \xrightarrow[k\to\infty]{} \nu,$$

for the weak topology (convergence against continuous functions) on $\mathcal{X}$.

If $(\mathcal{X}_k, d_k, \nu_k)$ and $(\mathcal{X}, d, \nu)$ are Polish spaces, not necessarily compact, equipped with σ-finite measures, then it is said that $\mathcal{X}_k$ converges to $\mathcal{X}$ in the local measured Gromov–Hausdorff topology if there are increasing sequences of compact sets $(K_k^{(\ell)})_{\ell\in\mathbb{N}}$ for each k, and $(K^{(\ell)})_{\ell\in\mathbb{N}}$, such that for each ℓ, the space $K_k^{(\ell)}$, seen as a subspace of $\mathcal{X}_k$, converges in the measured Gromov–Hausdorff topology to $K^{(\ell)}$; and the union of all $K^{(\ell)}$ is dense in $\mathcal{X}$.

If the spaces $(\mathcal{X}_k, d_k, \nu_k, \star_k)$ and $(\mathcal{X}, d, \nu, \star)$ are locally compact pointed Polish spaces equipped with locally finite measures, then it is said that $\mathcal{X}_k$ converges to $\mathcal{X}$ in the pointed measured Gromov–Hausdorff topology if there are sequences $R_k \to \infty$ and $\varepsilon_k \to 0$, and measurable pointed ε_k-isometries $B(\star_k, R_k]) \to B(\star, R_k])$, such that

$$(f_k)_\# \nu_k \xrightarrow[k\to\infty]{} \nu,$$

where the convergence is now the weak-$*$ topology of convergence against compactly supported continuous functions.

Remark 27.30. As already remarked for the plain Gromov–Hausdorff topology, one might also require in the definition of pointed measured Gromov–Hausdorff topology that $R_k = (\varepsilon_k)^{-1}$, but I shall not do so here.

From the material in this chapter it is easy to derive the following compactness criterion:

Proposition 27.31 (compactness in measured Gromov–Hausdorff topology). *(i) Let $R > 0$, $D > 0$, and $0 < m \leq M$ be finite positive constants, and let $\mathcal{F}$ be a family of compact metric-measure spaces, such that for each $(\mathcal{X}, d, \nu) \in \mathcal{F}$ the diameter of $(\mathcal{X}, d)$ is bounded above by $2R$, and the measure ν has a doubling constant bounded above by D, and a total mass $\nu[\mathcal{X}]$ bounded above by M and bounded below by m. Then $\mathcal{F}$ is precompact in the measured Gromov–Hausdorff topology. In particular, any weak cluster space $(\mathcal{X}_\infty, d_\infty, \nu_\infty)$ satisfies $\mathrm{Spt}\, \nu_\infty = \mathcal{X}_\infty$.*

(ii) Let $\mathcal{F}$ be a family of locally compact pointed Polish measure spaces. Assume that for each R, there is a constant $D = D(R)$ such that for each $(\mathcal{X}, d, \nu, \star) \in \mathcal{F}$ the measure ν is D-doubling on $B_R(\star)$. Further assume the existence of $m, M > 0$ such that $m \leq \nu[B_1(\mathcal{X})] \leq M$ for all $(\mathcal{X}, d, \nu) \in \mathcal{F}$. Then $\mathcal{F}$ is precompact in the pointed measured Gromov–Hausdorff topology. In particular, any weak cluster space $(\mathcal{X}_\infty, d_\infty, \nu_\infty)$ satisfies $\mathrm{Spt}\, \nu_\infty = \mathcal{X}_\infty$.

Proof. Part (i) follows from the combination of Proposition 27.28, Theorem 27.10 and Proposition 27.25. Part (ii) follows in addition from the definition of pointed measured Gromov–Hausdorff convergence and Proposition 27.23. Note that in (ii), the doubling assumption is used to prevent the formation of "spikes", but also to ensure uniform bounds on the mass of balls of radius R for *any* R, once it is known for *some* R. (Here I chose $R = 1$, but of course any other choice would have done.) □

Bibliographical Notes

It is well-known to specialists, but not necessarily obvious to others, that the Prokhorov distance, as defined in (6.5), coincides with the expression given in the beginning of this chapter; see for instance [365, Remark 1.29].

My presentation of the Gromov–Hausdorff topology mainly followed the very pedagogical book of Burago, Burago and Ivanov [81]. In particular, one can find there the proofs of Theorems 27.9 and 27.10. Other classical sources about the convergence of metric spaces and measure-metric spaces are the book by Gromov [196], Petersen [**?**, ChapterXXXX], and the survey paper by Fukaya [172]. The reader can also consult S. Evans' presentation in his own Saint-Flour lecture notes. I shall make further comments on [81] below.

Information about Mr. Magoo, the famous cartoon character, can be found on the Web site `http://www.toontracker.com`.

The definition of a metric cone can be found in [81, Section 3.6.2], and the notion of tangent cone is explained in [81, p. 321]. There are actually two definitions of tangent cone: either as the metric cone over the space of directions, or as the Gromov–Hausdorff limit of rescalings; read carefully Remarks 9.1.40 to 9.1.42 in [81] to avoid traps (or see [272]). For Alexandrov spaces with curvature bounded below, both notions coincide, see [81, Section 10.9] and the references therein.

Proposition 27.20 appears, in a form which is not exactly the one that I stated, but quite close to, in Gromov [194, p. 66] and [200, Appendix A]. The reader should have no difficulty in adapting the statements there into Proposition 27.20 (or re-do the proof of

the Ascoli theorem). Proposition 27.22 is rather easy to prove, and anyway in the next chapter we shall prove some more complicated theorems.

A classical source for the measured Gromov–Hausdorff topology is Gromov's book [196]. The point of view mainly used there consists in forgetting the Gromov–Hausdorff distance and "using the measure to kill infinity"; so the distances that are found there would be of the sort of d_{MGW_p} or d_{MGP}. Sturm made a detailed study of d_{MGW_2} in [335] and advocated it as a natural distance in the context of optimal transport.

The alternative point of view in which one takes care of both the metric and the measure was introduced by Fukaya [171]. This is the one with which Lott and myself used in our study of displacement convexity in length spaces [247].

The pointed Gromov–Hausdorff topology is presented for instance in [81]; it has become very popular as a way to study tangent spaces in the absence of smoothness.

In the context of optimal transport, the pointed Gromov–Hausdorff topology was used independently in [15, Section 12.4] and in [247, Appendix A]. Also in a context of optimal transport, the pointed measured Gromov–Hausdorff topology was used in [247, Appendix F].

The fact that a locally compact length space is automatically boundedly compact is part of the generalized version of the Hopf-Rinow theorem for length spaces [81, Theorem 2.5.28].

I introduced the definition of *local* Gromov–Hausdorff topology for the purpose of these notes; it looks to me quite natural if one wants to preserve the idea of pointing in a setting that might not necessarily be locally compact. This is not such a serious issue and the reader who does not like this notion can still go on with the rest of these notes. Still, let me advocate for it as a natural concept to treat the Gromov–Hausdorff convergence on the Wasserstein space over a noncompact metric space (see Chapter 28).

The statement of completeness of the Gromov–Hausdorff space appears in Gromov's book [196, p. 78]. Its proof can be found e.g. in Fukaya [172, Theorem 1.5], or in the book by Petersen [**?**, Proposition 1.7].

The theorem briefly alluded to in Remark 27.8 states the following: If M is an n-dimensional compact manifold, and $(M_k)_{k\in\mathbb{N}}$ is a sequence of n-dimensional compact manifolds converging to M, with uniform upper and lower bounds on the sectional curvatures, and a volume which is uniformly bounded below, then M_k is diffeomorphic to M for k large enough. This result is due to Gromov (after precursors by Shikata); see e.g. [172, Chapter 3] for a proof and references.

As a last remark, the construction of approximate isometries from correspondences, as performed in [81], uses the full axiom of choice (on p. 258: "For each x, choose $f(x) \in Y$"). So I should sketch a proof which does not use it. Let $\mathcal{R}$ be a correspondence with distortion η, the problem is to construct an ε-isometry $f : \mathcal{X} \to \mathcal{Y}$ for any $\varepsilon > \eta$. Let $\mathcal{D}$ be a countable dense subset in $\mathcal{X}$. Choose δ so small that $2\delta + \eta < \varepsilon$. Cover $\mathcal{X}$ by finitely many disjoint sets A_k, such that each A_k is included in some ball $B(x_k, \delta)$, with $x_k \in \mathcal{D}$. Then for all $x \in \mathcal{D}$ choose $f(x)$ in relation with x. (This only uses the countable version of the axiom of choice.) Finally, for each $x \in A_k$ define $f(x) = f(x_k)$. It is easy to check that $\mathrm{dis}\,(f) \le 2\delta + \eta < \varepsilon$.

(This axiomatic issue is also the reason why I work with approximated inverses that are (4ε)-isometries rather than (3ε)-isometries....)

28

Stability of optimal transport

This chapter is devoted to the following issue: Consider a family of length spaces $\mathcal{X}_k$ which converges to a length space $\mathcal{X}$, does it follow that certain basic objects in the theory of optimal transport on $\mathcal{X}_k$ "pass to the limit"? In this chapter I shall show that the answer is affirmative: one of the main results is that the Wasserstein space $P_2(\mathcal{X}_k)$ converges, in Gromov–Hausdorff sense, to the Wasserstein space $P_2(\mathcal{X})$. Then I shall consider the stability of *dynamical* optimal transference plans, and related objects (displacement interpolation, kinetic energy, etc.) Compact spaces will be considered first, and will be the basis for the subsequent treatment of noncompact spaces.

Optimal transport in a nonsmooth setting

Most of the objects that were introduced and studied in the context of optimal transport on Riemannian manifolds still make sense on a general metric-measure length space $(\mathcal{X}, d, \nu)$, satisfying certain regularity assumptions. I shall assume here that $(\mathcal{X}, d)$ is a **locally compact, complete separable geodesic length space** equipped with a σ-finite reference Borel measure ν.

From general properties of such spaces, plus the results in Chapters 6 and 7, it follows that

- the cost function $c(x, y) = d(x, y)^2$ is associated with the coercive Lagrangian action $\mathcal{A}(\gamma) = \mathcal{L}(\gamma)^2$, and minimizers are constant-speed, minimizing geodesics, the collection of which is denoted by $\Gamma(\mathcal{X})$;

- for any given μ_0, μ_1 in $P_2(\mathcal{X})$, the optimal total cost $C(\mu_0, \mu_1)$ is finite and there exists at least one optimal transference plan $\pi \in P(\mathcal{X} \times \mathcal{X})$ with marginals μ_0 and μ_1;

- the 2-Wasserstein space $P_2(\mathcal{X})$, equipped with the 2-Wasserstein distance, is a complete separable geodesic space;

- a displacement interpolation $(\mu_t)_{0 \le t \le 1}$ can be defined either as a geodesic in $P_2(\mathcal{X})$, or as $(e_t)_\# \Pi$, where e_t is the evaluation at time t, and Π is a dynamical optimal transference plan, i.e. the law of a random geodesic whose endpoints form an optimal coupling of (μ_0, μ_1).

One can also introduce the interpolant density $\rho_t = \rho_t(x)$ as the density of μ_t with respect to the reference measure ν.

Many of the statements that were available in the Riemannian setting can be recast in terms of these objects. An importance difference however is the absence of any "explicit" description of optimal couplings in terms of $d^2/2$-convex maps ψ. So expressions involving $\nabla\psi$ will not a priori make sense, unless we find an intrinsic reformulation in terms of the above-mentioned objects. For instance,

$$\int \rho_0(x)|\nabla\psi(x)|^2\,d\nu(x) = \int d\Big(x, \exp_x \nabla\psi(x)\Big)^2\,d\mu_0(x) = W_2(\mu_0,\mu_1)^2. \tag{28.1}$$

There is a more precise procedure which allows one to make sense of $|\nabla\psi|$, even if $\nabla\psi$ itself does not. The crucial observation, as in (28.1), is that $|\nabla\psi(x)|$ can be identified with the *length* $\mathcal{L}(\gamma)$ of the geodesic γ joining $\gamma(0) = x$ to $\gamma(1) = y$. In the next paragraph I shal develop this remark. The hasty reader can skip this bit and go directly to the section about the convergence of Wasserstein spaces.

Kinetic energy and speed

Definition 28.1 (kinetic energy). *Let $\mathcal{X}$ be a locally compact Polish length space, and let $\Pi \in P(\Gamma(\mathcal{X}))$ be a dynamical transference plan. For each $t \in (0,1)$ define the associated kinetic energy $\varepsilon_t(dx)$ by the formula*

$$\varepsilon_t = (e_t)_\# \left(\frac{\mathcal{L}^2}{2}\,\Pi\right).$$

If ε_t is absolutely continuous with respect to μ_t, define the speed field $|v|(t,x)$ by the formula

$$|v|(t,x) = \sqrt{2\,\frac{d\varepsilon_t}{d\mu_t}}.$$

Remark 28.2. If $\mathcal{X}$ is bounded, then $\varepsilon_t \le C\mu_t$, where $C = (\text{diam}\ \mathcal{X})^2/2$; then $|v|$ is well-defined (up to modification on a set of zero μ_t-measure), and bounded almost surely by $\sqrt{2C} = \text{diam}\,(\mathcal{X})$.

Remark 28.3. If γ is a geodesic curve, then $\mathcal{L}(\gamma) = |\dot\gamma|(t)$, whatever $t \in (0,1)$. Assume that $\mathcal{X}$ is a Riemannian manifold M, and geodesics in the support of Π do not cross at intermediate times. (As we know from Chapter 8, this is the case if Π is an *optimal* dynamical transference plan.) Then for each $t \in (0,1)$ and $x \in M$ there is at most one geodesic $\gamma = \gamma^{x,t}$ such that $\gamma(t) = x$. So

$$\varepsilon_t(dx) = |\dot\gamma^{\,x,t}(t)|^2\,[(e_t)_\#\Pi](dx) = |\dot\gamma|^2\,\mu_t(dx);$$

then $|v|(t,x)$ really is $|\dot\gamma^{x,t}|$, that is, the speed at time t and position x. So Definition 28.1 is consistent with the usual notions of kinetic energy and speed field (speed = norm of the velocity).

Particular Case 28.4. Let M be a Riemannian manifold, and let μ_0, μ_1 be two probability measures in $P_2(M)$, μ_0 being absolutely continuous with respect to Lebesgue measure. Let ψ be a $d^2/2$-convex function such that $\exp(\nabla\psi)$ is the optimal transport map from μ_0 to μ_1, and let ψ_t be obtained by solving the forward Hamilton–Jacobi equation $\partial_t\psi_t + |\nabla\psi_t|^2/2 = 0$ starting from the initial data $\psi_0 = \psi$. Then the speed $|v|(t,x)$ coincides, μ_t-almost surely, with $|\nabla\psi_t(x)|$.

The kinetic energy is a nonnegative measure, while the field speed is a function. Both objects will enjoy good stability properties under Gromov–Hausdorff approximation. But under adequate assumptions, the velocity field will also enjoy compactness properties in the *uniform* topology. This comes from the next statement.

Theorem 28.5 (regularity of the speed field). *Let $(\mathcal{X}, d)$ be a compact length space, let $\Pi \in P(\Gamma(\mathcal{X}))$ be a dynamical optimal transference plan, let $(\mu_t)_{0\le t\le 1}$ be the associated displacement interpolation, and $|v| = |v|(t,x)$ the associated speed field. Then, for each $t \in (0,1)$ one can modify $|v|(t,\cdot)$ on a μ_t-negligible set in such a way that for* all *$x, y \in \mathcal{X}$,*

$$\Big|\, |v|(t,x) \,-\, |v|(t,y) \,\Big| \;\le\; \frac{C\sqrt{\mathrm{diam}\ \mathcal{X}}}{\sqrt{t(1-t)}}\ \sqrt{d(x,y)}, \tag{28.2}$$

where C is a numeric constant. In particular, $|v|(t,\cdot)$ is Hölder-1/2.

Proof of Theorem 28.5. Let t be a fixed time in $(0,1)$. Let γ_1 and γ_2 be two minimizing geodesics in the support of Π, and let $x = \gamma_1(t)$, $y = \gamma_2(t)$. Then, by Theorem 8.22,

$$\big|\mathcal{L}(\gamma_1) - \mathcal{L}(\gamma_2)\big| \le \frac{C\sqrt{\mathrm{diam}\,(\mathcal{X})}}{\sqrt{t(1-t)}}\ \sqrt{d(x,y)}. \tag{28.3}$$

Let $\mathcal{X}_t$ be the union of all $\gamma(t)$, for γ in the support of Π. For a given $x \in \mathcal{X}_t$, there might be several geodesics γ passing through x, but (as a special case of (28.3)) they will all have the same length; define $|v|(t,x)$ to be that length. This is a measurable function, since it can be rewritten

$$|v|(t,x) = \int_\Gamma \mathcal{L}(\gamma)\, \Pi(d\gamma|\gamma(t) = x),$$

where $\Pi(d\gamma|\gamma(t) = x)$ is of course the disintegration of Π with respect to μ_t, the law of γ_t. Then $|v|(t,\cdot)$ is an admissible density for ε_t, and as a consequence of (28.3) it satisfies (28.2) for all $x, y \in \mathcal{X}_t$.

Now to extend $|v|(t,x)$ on the whole of $\mathcal{X}$, I shall adapt a well-known extension theorem for Lipschitz functions defined on a subset of a metric space. Let $H := C\sqrt{\mathrm{diam}\ \mathcal{X}}/(t(1-t))$, so that $|v|$ is Hölder-1/2 with constant H on $\mathcal{X}_t$. Then define, for $x \in \mathcal{X}$,

$$w(x) := \inf_{y\in\mathcal{X}_t} \Big[H\sqrt{d(x,y)} + |v|(t,y) \Big].$$

It is clear that $w \ge 0$, and the estimate (28.2) easily implies that $w(x) = |v|(t,x)$ for any $x \in \mathcal{X}_t$. Next, whenever x and x' are two points in $\mathcal{X}$, one has

$$\begin{aligned}
w(x) - w(x') &= \inf_{y\in\mathcal{X}_t} \Big[H\sqrt{d(x,y)} + |v|(t,y) \Big] - \inf_{y'\in\mathcal{X}_t} \Big[H\sqrt{d(x',y')} + |v|(t,y') \Big] \\
&= \sup_{y'\in\mathcal{X}_t} \inf_{y\in\mathcal{X}_t} \Big[H\sqrt{d(x,y)} - H\sqrt{d(x',y')} + |v|(t,y) - |v|(t,y') \Big] \\
&\le H \sup_{y'\in\mathcal{X}_t} \inf_{y\in\mathcal{X}_t} \Big[\sqrt{d(x,y)} - \sqrt{d(x',y')} + \sqrt{d(y,y')} \Big] \\
&\le H \sup_{y'\in\mathcal{X}_t} \Big[\sqrt{d(x,y')} - \sqrt{d(x',y')} \Big].
\end{aligned} \tag{28.4}$$

But

$$\sqrt{d(x,y')} \le \sqrt{d(x,x') + d(x',y')} \le \sqrt{d(x,x')} + \sqrt{d(x',y')};$$

so (28.4) is bounded by $H\sqrt{d(x,x')}$.

To summarize: w coincides with $|v|(t,\cdot)$ on $\mathcal{X}_t$, and it satisfies the same Hölder-1/2 estimate. Since μ_t is concentrated on $\mathcal{X}_t$, w is also an admissible density for ε_t, so we can take it as the new definition of $|v|(t,\cdot)$, and then (28.2) holds true on the whole of $\mathcal{X}$. □

Convergence of the Wasserstein space

The main goal of this section is the proof of the convergence of the Wasserstein space $P_2(\mathcal{X})$, as expressed in the next statement.

Theorem 28.6. *Let $(\mathcal{X}_k)_{k\in\mathbb{N}}$ and $\mathcal{X}$ be compact metric spaces such that*

$$\mathcal{X}_k \underset{GH}{\longrightarrow} \mathcal{X}.$$

Then

$$P_2(\mathcal{X}_k) \underset{GH}{\longrightarrow} P_2(\mathcal{X}).$$

Theorem 28.6 will come as an immediate corollary of the following more precise results:

Proposition 28.7. *Let $f : (\mathcal{X}_1, d_1) \to (\mathcal{X}_2, d_2)$ be an ε-isometry between two metric spaces. Then the map $f_\#$, defined by $f_\#(\mu) = f_\#\mu$, is an $\widetilde{\varepsilon}$-isometry between $P_2(\mathcal{X}_1)$ and $P_2(\mathcal{X}_2)$, where*

$$\widetilde{\varepsilon} = 6\,\varepsilon + 2\sqrt{\varepsilon\,(2\,\mathrm{diam}\,(\mathcal{X}_2) \,+\, 2\varepsilon)} \;\le 10\big(\varepsilon + \sqrt{\varepsilon\ \mathrm{diam}\,(\mathcal{X}_2)}\big). \tag{28.5}$$

Remark 28.8. The map $f_\#$ is continuous if and only if f itself is continuous (which in general is not the case).

Proof of Proposition 28.7. Let us start with the proof of (28.5). Let f be an ε-isometry, and let f' be an ε-inverse for f. Recall that f' is a (4ε)-isometry and satisfies (27.4).

Given μ_1 and μ_1' in $P_2(\mathcal{X}_1)$, let π_1 be an optimal transference plan between μ_1 and μ_1'. Define

$$\pi_2 := \big(f, f\big)_\# \pi_1.$$

Obviously, π_2 is a transference plan between $f_\#\mu_1$ and $f\#\mu_1'$. Then

$$W_2\big(f_\#\mu_1, f_\#\mu_1'\big)^2 \le \int_{\mathcal{X}_2\times\mathcal{X}_2} d_2(x_2,y_2)^2\, d\pi_2(x_2,y_2) = \int_{\mathcal{X}_1\times\mathcal{X}_1} d_2\big(f(x_1),f(y_1)\big)^2\, d\pi_1(x_1,y_1). \tag{28.6}$$

As

$$\begin{aligned}\Big|d_2(f(x_1),f(y_1))^2 - d_1(x_1,y_1)^2\Big| \\ = \Big|d_2(f(x_1),f(y_1)) - d_1(x_1,y_1)\Big|\,\Big(d_2(f(x_1),f(y_1)) + d_1(x_1,y_1)\Big)\end{aligned}$$

we have

$$\Big|d_2(f(x_1),f(y_1))^2 - d_1(x_1,y_1)^2\Big| \;\le\; \varepsilon\,\big(\mathrm{diam}\,(\mathcal{X}_1) + \mathrm{diam}\,(\mathcal{X}_2)\big). \tag{28.7}$$

Therefore

$$W_2\big(f_\#\mu_1, f_\#\mu_1'\big)^2 \le W_2(\mu_1,\mu_1')^2 + \varepsilon\,\big(\mathrm{diam}\,(\mathcal{X}_1) + \mathrm{diam}\,(\mathcal{X}_2)\big), \tag{28.8}$$

hence

$$W_2\big(f_\#\mu_1, f_\#\mu_1'\big) \le W_2(\mu_1,\mu_1') + \sqrt{\varepsilon\,\big(\mathrm{diam}\,(\mathcal{X}_1) + \mathrm{diam}\,(\mathcal{X}_2)\big)}. \tag{28.9}$$

It follows from the definition of an ε-isometry that $\mathrm{diam}\,(\mathcal{X}_1) \le \mathrm{diam}\,(\mathcal{X}_2) + \varepsilon$; so (28.9) leads to

$$W_2\big(f_\#\mu_1, f_\#\mu_1'\big) \le W_2(\mu_1,\mu_1') + \sqrt{\varepsilon\,\big(2\,\mathrm{diam}\,(\mathcal{X}_2) + \varepsilon\big)}, \tag{28.10}$$

which shows that $f_\#$ does not increase distances much.

On the other hand, $\mathrm{diam}\,(\mathcal{X}_2) \le \mathrm{diam}\,(\mathcal{X}_1) + 2\varepsilon$; so (28.9) also implies

$$W_2\big(f_\#\mu_1, f_\#\mu_1'\big) \le W_2(\mu_1,\mu_1') + \sqrt{\varepsilon\,\big(2\,\mathrm{diam}\,(\mathcal{X}_1) + 2\varepsilon\big)}. \tag{28.11}$$

Exchanging the roles of $\mathcal{X}_1$ and $\mathcal{X}_2$, and applying (28.8) to the map f' and the measures $f_\#\mu_1$ and $f_\#\mu_1'$, we obtain

$$W_2\Big((f')_\#(f_\#\mu_1), (f')_\#(f_\#\mu_1')\Big) \le W_2\big(f_\#\mu_1, f_\#\mu_1'\big) + \sqrt{4\varepsilon\,(2\,\mathrm{diam}\,(\mathcal{X}_2) + 8\varepsilon)}. \tag{28.12}$$

(The factor 4 is because f' is a (4ε)-isometry.) Since $f'\circ f$ is an admissible Monge transport between μ_1 and $(f'\circ f)_\#\mu_1$, or between μ_1' and $(f'\circ f)_\#\mu_1'$, which moves points by a distance at most 3ε, we have

$$W_2\Big((f'\circ f)_\#\mu_1, \mu_1\Big) \le 3\varepsilon; \qquad W_2\Big((f'\circ f)_\#\mu_1', \mu_1'\Big) \le 3\varepsilon. \tag{28.13}$$

Then by (28.12) and the triangle inequality,

$$\begin{aligned} W_2(\mu_1,\mu_1') &\le W_2(\mu_1, (f'\circ f)_\#\mu_1) + W_2\big((f'\circ f)_\#\mu_1, (f'\circ f)_\#\mu_1'\big) + W_2((f'\circ f)_\#\mu_1', \mu_1') \\ &\le 3\varepsilon + W_2\big(f_\#\mu_1, f_\#\mu_1'\big) + \sqrt{4\varepsilon(2\,\mathrm{diam}\,(\mathcal{X}_2) + 8\varepsilon)} + 3\varepsilon. \end{aligned} \tag{28.14}$$

Equations (28.10) and (28.14) together show that $f_\#$ distorts distances by at most $\widetilde{\varepsilon}$.

It remains to show that $f_\#$ is approximately surjective. To do this, pick up some $\mu_2 \in P_2(\mathcal{X}_2)$, and consider the Monge transport $f\circ f'$ from μ_2 to $(f\circ f')_\#\mu_2$. Since $f\circ f'$ moves points by a distance at most ε, we have $W_2(\mu_2, f_\#(f'_\#\mu_2)) \le \varepsilon$. This concludes the proof that $f_\#$ is an $\widetilde{\varepsilon}$-isometry. □

Compactness of dynamical transference plans and related objects

The issue now is to show that dynamical transference plans enjoy good stability properties in a Gromov–Hausdorff approximation. The main technical difficulty comes from the fact that ε-isometries, being in general discontinuous, will not map geodesic paths into continuous paths. So we shall be led to work on the horribly large space of *measurable* paths $[0,1] \to \mathcal{X}$. I shall daringly imbed this space in the even much larger space of probability measures on $[0,1]\times\mathcal{X}$, via the identification $\gamma \longmapsto \overline{\gamma}$, where

$$\overline{\gamma} = (\mathrm{Id}\,, \gamma)_\#\lambda, \tag{28.15}$$

and λ is the Lebesgue measure on $[0,1]$. In loose notation,

$$\overline{\gamma}(dt\,dx) = \delta_{x=\gamma(t)}\,dt. \tag{28.16}$$

Of course, the first marginal of such a measure is always the Lebesgue measure. That is, if $\tau : [0,1]\times\mathcal{X} \to [0,1]$ is the projection on the first factor then $\tau_\#\overline{\gamma} = \lambda$. Moreover, the uniqueness of conditional measures shows that if $\overline{\gamma_1} = \overline{\gamma_2}$, then $\gamma_1 = \gamma_2$ λ-almost surely, and therefore actually $\gamma_1 = \gamma_2$ (because γ_1, γ_2 are continuous).

In this way there is an injection $i : \mathit{\Gamma} \to P([0,1]\times\mathcal{X})$, defined by $i(\gamma) = \overline{\gamma}$, which can be thought of as an "inclusion". So any $\mathit{\Pi} \in P(\mathit{\Gamma})$ can be identified with its push-forward $i_\#\mathit{\Pi} \in P(P([0,1]\times\mathcal{X}))$. This point of view is reminiscent of the theory of Young measures; one of its advantages is that the space $P([0,1]\times\mathcal{X})$ is separable, while the space of measurable paths with values in $\mathcal{X}$ is not.

The next theorem expresses the stability of the main objects associated with transport (optimal or not).

Theorem 28.9 (stability of transport under Gromov–Hausdorff convergence). *Let $(\mathcal{X}_k, d_k)_{k\in\mathbb{N}}$ and $(\mathcal{X}, d)$ be compact length spaces such that $\mathcal{X}_k$ converges in the Gromov–Hausdorff topology to $\mathcal{X}$. Let $f_k : \mathcal{X}_k \to \mathcal{X}$ be ε_k-isometries, with $\varepsilon_k \to 0$. For each k, let Π_k be a Borel probability measure on $\Gamma(\mathcal{X}_k)$; let further $\pi_k = (e_0, e_1)_\# \Pi_k$ and $\mu_{k,t} = (e_t)_\# \Pi_k$ be the transference plan and the measure-valued path associated with Π_k. Then, after extraction of a subsequence, still denoted with the index k for simplicity, there is a dynamical transference plan Π on $\mathcal{X}$, with associated transference plan π, and measure-valued path $(\mu_t)_{0\le t\le 1}$, such that*

(i) $\lim_{k\to\infty} (f_k\circ)_\# \Pi_k = \Pi$ *in the weak topology on* $P(P([0,1]\times\mathcal{X}))$;

(ii) $\lim_{k\to\infty} (f_k, f_k)_\# \pi_k = \pi$ *in the weak topology on* $P(\mathcal{X}\times\mathcal{X})$;

(iii) $\lim_{k\to\infty} (f_k)_\# \mu_{k,t} = \mu_t$ *in* $P_2(\mathcal{X})$ *uniformly in* t; *i.e.* $\lim_{k\to\infty} \sup_{t\in[0,1]} W_2(\mu_{k,t}, \mu_t) = 0$;

(iv) $\lim_{k\to\infty} (f_k)_\# \varepsilon_{k,t} = \varepsilon_t$, *in the weak topology of measures, for each* $t \in (0,1)$;

Assume further that each Π_k is an optimal dynamical transference plan, for the square distance cost function; then

(v) for each $t \in (0,1)$, there is a choice of the speed fields $|v_k|$ such that $\lim_{k\to\infty} |v_k| \circ f_k' = |v|$, in the uniform topology.

(vi) the limit Π is an optimal dynamical transference plan, so π is an optimal transference plan and $(\mu_t)_{0\le t\le 1}$ is a displacement interpolation. Furthermore,

Remark 28.10. Here $f_k\circ$ is the map $P([0,1]\times\mathcal{X}_k) \to P([0,1]\times\mathcal{X})$ arising from f_k. For example, $f_k\circ$ takes the Dirac mass supported on a single geodesic γ_k, to the Dirac mass supported on the path $f_k\circ\gamma_k : [0,1] \to \mathcal{X}$. Note that the latter path may be discontinuous.

Proof of Theorem 28.9. The proof is quite technical, so the reader might skip it at first reading and go directly to the last section of this chapter. In a first step, I shall establish the compactness of the relevant objects, and in a second step pass to the limit.

It will be convenient to regularize rough paths with the help of some continuous mollifiers. For $\delta \in (0, 1/2)$, define

$$\varphi^\delta(s) = \frac{\delta + s}{\delta^2}\, 1_{-\delta\le s<0} + \frac{\delta - s}{\delta^2}\, 1_{0<s\le\delta} \tag{28.17}$$

and

$$\varphi^\delta_+(s) = \varphi^\delta(s-\delta), \qquad \varphi^\delta_-(s) = \varphi^\delta(s+\delta). \tag{28.18}$$

Then $\mathrm{supp}\,\varphi^\delta_+ \subset [0, 2\delta]$ and $\mathrm{supp}\,\varphi^\delta_- \subset [-2\delta, 0]$. These functions have a graph that looks like a sharp "tent hat"; they all have unit integral on the real line and as $\delta \to 0$ they converge in the weak topology to the Dirac mass δ_0 at the origin.

Step 1: Compactness. First, $[0,1]\times\mathcal{X}$ is a compact metric space, so the same holds true for $P([0,1]\times\mathcal{X})$ and $P(P([0,1]\times\mathcal{X}))$. Hence, after extraction of a subsequence, the sequence $((f_k\circ)_\#\Pi_k)_{k\in\mathbb{N}}$ converges to some $\Pi \in P(P([0,1]\times\mathcal{X}))$. Taking a further subsequence, we can assume that $\lim_{k\to\infty}(f_k, f_k)_\#\pi_k = \pi$ for some $\pi \in P(\mathcal{X}\times\mathcal{X})$.

Next, since $\mathcal{X}$ is bounded and $\mathcal{X}_k \to \mathcal{X}$, there is a uniform bound C on the diameters $\mathrm{diam}\,(\mathcal{X}_k)$. So the lengths of all geodesics $\gamma_k \in \Gamma(\mathcal{X}_k)$ are all bounded by C, and $d(\gamma_k(s), \gamma_k(t)) \le C|s-t|$ for all times $s, t \in [0,1]$. It follows that $W_2(\mu_{k,s}, \mu_{k,t}) \le C|s-t|$, as $(e_s, e_t)_\#\Pi_k$ is a particular transference plan between $\mu_{k,s}$ and $\mu_{k,t}$. This shows that the paths $(\mu_{k,t})_{t\in[0,1],k\in\mathbb{N}}$ are uniformly continuous in t, with a uniform modulus of continuity. On the other hand, by Theorem 28.6, $(P_2(\mathcal{X}_k))_{k\in\mathbb{N}}$ converges in the Gromov–Hausdorff topology to $P_2(\mathcal{X})$. By Proposition 27.20, there is a subsequence (in k) of the

family $((f_k)_\# \mu_{k,t})_{t\in[0,1],k\in\mathbb{N}}$ which converges uniformly to a continuous curve $(\mu_t)_{t\in[0,1]} \in C([0,1]; P(\mathcal{X}))$.

Next, for each $t \in (0,1)$ the total mass of the measures $\varepsilon_{k,t}$ are bounded by $\mathrm{diam}\,(\mathcal{X}_k)^2/2 \le C^2/2$; so the same holds true for the measures $(f_k)_\#\varepsilon_{k,t}$, which therefore constitute a precompact family in the space of nonnegative measures. So up to extraction, we may assume that $(f_k)_\#\varepsilon_{k,t}$ converges weakly to some measure ε_t.

To conclude the proof of (i)–(iv), we should establish that
(a) Π is actually concentrated on $\Gamma(\mathcal{X})$;
(b) $\pi = (e_0, e_1)_\# \Pi$;
(c) $\mu_t = (e_t)_\# \Pi$;
(d) $\varepsilon_t = (e_t)_\# (\mathcal{L}^2 \Pi)/2$.

Step 2: Embedding in probability measures and passing to the limit

I shall first mollify the geodesic-defining condition "$\mathcal{L}(\gamma) = d(\gamma(0), \gamma(1))$" in such a way that the resulting condition will pass to the limit under weak convergence of probability measures. Given $\delta \in (0, 1/2)$, a continuous path $\gamma : [0,1] \to \mathcal{X}$, and times t_0, s_0 with $0 \le t_0 < s_0 \le 1$, define

$$\mathcal{L}^\delta_{t_0 \to s_0}(\gamma) = \int_0^1 \int_0^1 d(\gamma(t), \gamma(s))\, \varphi^\delta_+(t - t_0)\, \varphi^\delta_-(s - s_0)\, dt\, ds. \tag{28.19}$$

This function extends into a continuous function on $P([0,1] \times \mathcal{X})$, still denoted $\mathcal{L}^\delta_{t_0\to s_0}$ for simplicity, by

$$\mathcal{L}^\delta_{t_0 \to s_0}(\sigma) = \int_{[0,1]\times\mathcal{X}} \int_{[0,1]\times\mathcal{X}} d(x,y)\, \varphi^\delta_+(t - t_0)\, \varphi^\delta_-(s - s_0)\, d\sigma(t,x)\, d\sigma(s,y). \tag{28.20}$$

Since f_k is an ε_k-approximation, if γ_k is a geodesic in $\mathcal{X}_k$ then

$$\begin{aligned}
\mathcal{L}^\delta_{t_0\to s_0}(f_k \circ \overline{\gamma_k}) &= \int_0^1 \int_0^1 \Big[d\big(\gamma_k(t), \gamma_k(s)\big) + O(\varepsilon_k) \Big]\, \varphi^\delta_+(t - t_0)\, \varphi^\delta_-(s - s_0)\, dt\, ds \qquad (28.21)\\
&= \int_0^1 \int_0^1 d\big(\gamma_k(t), \gamma_k(s)\big)\, \varphi^\delta_+(t - t_0)\, \varphi^\delta_-(s - s_0)\, dt\, ds + O(\varepsilon_k)\\
&= d\big(\gamma_k(0), \gamma_k(1)\big) \int_0^1 \int_0^1 |s - t|\, \varphi^\delta_+(t - t_0)\, \varphi^\delta_-(s - s_0)\, dt\, ds + O(\varepsilon_k)\\
&= d\big(\gamma_k(0), \gamma_k(1)\big) \Big(|s_0 - t_0| + O(\delta) \Big) \; + O(\varepsilon_k).
\end{aligned}$$

In particular, taking $t_0 = 0$ and $s_0 = 1$ gives

$$\mathcal{L}^\delta_{0\to 1}(f_k \circ \overline{\gamma_k}) = d\big(\gamma_k(0), \gamma_k(1)\big)\, (1 + O(\delta)) \; + O(\varepsilon_k). \tag{28.22}$$

Since all of the lengths $d(\gamma_k(0), \gamma_k(1))$ are uniformly bounded, we conclude that there is a constant C such that for all t_0 and s_0 as above,

$$\begin{cases} \Big| \mathcal{L}^\delta_{t_0\to s_0}(f_k \circ \overline{\gamma_k}) - |s_0 - t_0|\, \mathcal{L}^\delta_{0\to 1}(f_k \circ \overline{\gamma_k}) \Big| \;\le\; C(\delta + \varepsilon_k); \\ \\ \mathcal{L}^\delta_{0\to 1}(f_k \circ \overline{\gamma_k}) \le C. \end{cases} \tag{28.23}$$

Now for $\varepsilon, \delta > 0$, define

$$\Gamma_{\varepsilon,\delta}(\mathcal{X}) = \Big\{\sigma \in P([0,1]\times\mathcal{X});\quad \tau_{\#}\sigma = \lambda;$$
$$\Big|\mathcal{L}^{\delta}_{t_0\to s_0}(\sigma) - |s_0 - t_0|\,\mathcal{L}^{\delta}_{0\to 1}(\sigma)\Big| \le C(\delta+\varepsilon);\quad \mathcal{L}^{\delta}_{0\to 1}(\sigma) \le C\Big\}. \tag{28.24}$$

It is easy to see that $\Gamma_{\varepsilon,\delta}(\mathcal{X})$ is closed in $P([0,1]\times\mathcal{X})$. Moreover, for k large enough one has $\varepsilon_k \le \varepsilon$ and then $f_k \circ \overline{\gamma_k} \in \Gamma_{\varepsilon,\delta}(\mathcal{X})$ for any geodesic $\gamma_k \in \Gamma(\mathcal{X}_k)$. It follows that $(f_k\circ)_{\#}\Pi_k \in P(\Gamma_{\varepsilon,\delta}(\mathcal{X}))$ for k large enough; by passing to the limit, also $\Pi \in P(\Gamma_{\varepsilon,\delta}(\mathcal{X}))$. Since ε, δ are arbitrarily small,

$$\Pi \in \ P\Big(\bigcap_{\varepsilon,\delta>0} \Gamma_{\varepsilon,\delta}(\mathcal{X})\Big). \tag{28.25}$$

So to conclude the proof of (a) it suffices to prove

$$\bigcap_{\varepsilon,\delta>0} \Gamma_{\varepsilon,\delta}(\mathcal{X}) = \Gamma(\mathcal{X}). \tag{28.26}$$

So let $\sigma \in \bigcap_{\varepsilon,\delta>0} \Gamma_{\varepsilon,\delta}(\mathcal{X})$. Taking $\varepsilon \to 0$ in (28.24), we get

$$\Big|\mathcal{L}^{\delta}_{t_0\to s_0}(\sigma) - |s_0 - t_0|\,\mathcal{L}^{\delta}_{0\to 1}(\sigma)\Big| \ \le \delta. \tag{28.27}$$

In particular,

$$\mathcal{L}^{\delta}_{t_0\to s_0}(\sigma) \le C(|s_0 - t_0| + \delta). \tag{28.28}$$

In Lemma 28.11 below it will be shown that, as a consequence of (28.28), σ can be written as $(\mathrm{Id}\,,\gamma)_{\#}\lambda$ for some Lipschitz-continuous curve $\gamma : [0,1] \to \mathcal{X}$. Once that is known, the end of the proof of (a) is straightforward: Since

$$\mathcal{L}^{\delta}_{t_0\to s_0}(\sigma) = d\big(\gamma(t_0),\gamma(s_0)\big) + O(\delta), \tag{28.29}$$

the inequality (28.27) becomes, in the limit $\delta \to 0$,

$$d\big(\gamma(t_0),\gamma(s_0)\big) = |s_0 - t_0|\,d\big(\gamma(0),\gamma(1)\big). \tag{28.30}$$

This implies that $\mathcal{L}(\gamma) = d(\gamma(0),\gamma(1))$, so γ is a geodesic curve. This concludes the proof of (a), and of part (i) of Theorem 28.9 at the same time.

Now I shall use a similar reasoning for the convergence of the marginals of Π. Given $\Phi \in C(\mathcal{X}\times\mathcal{X})$ and $\gamma \in \Gamma(\mathcal{X})$, put

$$\Phi^{\delta}(\gamma) = \int_0^1 \int_0^1 \Phi\big(\gamma(t),\gamma(s)\big)\,\varphi^{\delta}_+(t)\,\varphi^{\delta}_-(s-1)\,dt\,ds. \tag{28.31}$$

As before, this extends to a continuous function on $P([0,1]\times\mathcal{X})$ by

$$\Phi^{\delta}(\sigma) = \int_{[0,1]\times\mathcal{X}} \int_{[0,1]\times\mathcal{X}} \Phi(x,y)\,\varphi^{\delta}_+(t)\,\varphi^{\delta}_-(s-1)\,d\sigma(t,x)\,d\sigma(s,y). \tag{28.32}$$

By part (i) of the theorem,

$$\int_{\Gamma(\mathcal{X}_k)} \Phi^{\delta}(f_k\circ\gamma_k)\,d\Pi_k(\gamma_k) \longrightarrow \int_{\Gamma(\mathcal{X})} \Phi^{\delta}(\gamma)\,d\Pi(\gamma). \tag{28.33}$$

Let us examine the behavior of the two sides of (28.33) as $\delta \to 0$. If γ is a geodesic on $\mathcal{X}$, the continuity of Φ and γ implies that $\Phi^\delta(\gamma) \longrightarrow \Phi(\gamma(0), \gamma(1))$ as $\delta \to 0$. Then by dominated convergence,

$$\begin{aligned} \int_{\Gamma(\mathcal{X})} \Phi^\delta(\gamma)\, d\Pi(\gamma) &\longrightarrow \int_{\Gamma(\mathcal{X})} \Phi\big(\gamma(0), \gamma(1)\big)\, d\Pi(\gamma) \\ &= \int \Phi\, d(e_0, e_1)_\# \Pi = \int \Phi\, d\pi. \end{aligned} \tag{28.34}$$

As for the left-hand-side of (28.33), things are not so immediate because $f_k \circ \gamma_k$ may be discontinuous. However, for $0 \le t \le 2\delta$ one has

$$d\big(f_k(\gamma_k(0)), f_k(\gamma_k(t))\big) = d_k\big(\gamma_k(0), \gamma_k(t)\big) + O(\varepsilon_k) = O(\delta + \varepsilon_k), \tag{28.35}$$

where the implicit constant in the right-hand side is independent of γ_k. Similarly, for $1 - 2\delta \le s \le 1$, one has

$$d\big(f_k(\gamma_k(s)), f_k(\gamma_k(1))\big) = O(\delta + \varepsilon_k). \tag{28.36}$$

Then it follows from the uniform continuity of Φ that

$$\sup_{\gamma_k \in \Gamma(\mathcal{X}_k),\, t \in [0, 2\delta],\, s \in [1-2\delta, 1]} \left| \Phi\Big(f_k(\gamma_k(t)),\, f_k(\gamma_k(s))\Big) - \Phi\Big(f_k(\gamma_k(0)),\, f_k(\gamma_k(1))\Big) \right| \to 0 \tag{28.37}$$

as $\delta \to 0$ and $k \to \infty$. So in this limit, the left-hand-side of (28.33) is well approximated by

$$\begin{aligned} \int_{\Gamma(\mathcal{X}_k)} \Phi\big(f_k(\gamma_k(0)), f_k(\gamma_k(1))\big)\, d\Pi_k(\gamma_k) &= \int_{\mathcal{X}\times\mathcal{X}} \Phi\, d\Big[(f_k, f_k)_\#\big((e_0, e_1)_\# \Pi_k\big)\Big] \\ &= \int_{\mathcal{X}\times\mathcal{X}} \Phi\, d(f_k, f_k)_\# \pi_k. \end{aligned} \tag{28.38}$$

The comparison of (28.33), (28.34) and (28.38) shows that $(f_k, f_k)_\# \pi_k$ converges to π, which concludes the proof of (b).

As for (c) we just have to show that $\lim_{k\to\infty} (f_k)_\# \mu_{k,t_0} = \mu_{t_0}$ for all $t_0 \in [0, 1]$. The argument is quite similar to the proof of (b). Assume, for example, that $t_0 < 1$. Given $\Phi \in C(\mathcal{X})$, define

$$\Phi^\delta_{t_0}(\gamma) = \int_0^1 \Phi(\gamma(t))\, \varphi^\delta_+(t - t_0)\, dt. \tag{28.39}$$

This extends to a continuous function on $P([0, 1] \times \mathcal{X})$, so

$$\int_{\Gamma(\mathcal{X}_k)} \Phi^\delta_{t_0}(f_k \circ \overline{\gamma_k})\, d\Pi_k(\gamma_k) \longrightarrow \int_{\Gamma(\mathcal{X})} \Phi^\delta_{t_0}(\gamma)\, d\Pi(\gamma). \tag{28.40}$$

The right-hand side converges to $\int_{\mathcal{X}} \Phi(x)\, d\mu_{t_0}(x)$ as $\delta \to 0$, while the left-hand side is well approximated by $\int_{\mathcal{X}} \Phi(f_k(x))\, d\mu_{k,t_0}(x)$. The conclusion follows.

The proof of (d) also follows a similar reasoning.

Let us finally turn to the proof of statements (v) and (vi) in the theorem. In the sequel, it will be assumed that each Π_k is an *optimal* dynamical optimal transference plan. In view of Theorem 28.5, for each $t \in (0, 1)$, the velocity fields $|v_{k,t}|$ can be chosen in such

a way that they satisfy a uniform Hölder-1/2 estimate. Then the precompactness of $|v_{k,t}|$ follows from Ascoli's theorem, in the form of Proposition 27.20. So up to extraction, we may assume that $|v_{k,t}| \circ f_k'$ converges uniformly to some function $|v_t|$. It remains to show that $|v_t|^2/2$ is an admissible density for ε, at each time $t \in (0,1)$. For simplicity I shall omit the time variable, so t is implicit and fixed in $(0,1)$. Since there is a uniform bound on the diameter of the spaces $\mathcal{X}_k$, the function $|v_k|^2 \circ f_k'$ converges uniformly to $|v|^2$. By uniform continuity of $|v_k|^2$, the difference between $|v_k|^2$ and $|v_k|^2 \circ (f_k' \circ f_k)$ is bounded by $\eta(k)$, where $\eta(k) \to 0$ as $k \to \infty$. After going back to the definitions of push-forward and weak convergence, it follows that

$$\lim_{k\to\infty} \big(|v_k|^2 \circ f_k'\big)\, (f_k)_\# \mu_k = \lim_{k\to\infty} (f_k)_\# \big(|v_k|^2\, \mu_k\big) = 2 \lim_{k\to\infty} (f_k)_\# \varepsilon_k = 2\,\varepsilon. \tag{28.41}$$

Since $|v_k|^2 \circ f_k'$ converges *uniformly* to $|v|^2$, and $(f_k)_\# \mu_k$ converges weakly to μ, the left-hand side in (28.41) converges weakly to $|v|^2\, \mu$. It follows that $|v|^2/2$ is an admissible density for the kinetic energy ε. This concludes the proof of (v).

The proof of (vi) is easy now. Since $\pi = \lim (f_k, f_k)_\# \pi_k$ and f_k is an approximate isometry,

$$\int d(x_0,x_1)^2\, d\pi(x_0,x_1) = \lim_{k\to\infty} \int d(f_k(x_0), f_k(x_1))^2\, d\pi_k(x_0,x_1) = \lim_{k\to\infty} \int d_k(x_0,x_1)^2\, d\pi_k(x_0,x_1). \tag{28.42}$$

By assumption, π_k is optimal for each k, so

$$\int d_k(x_0,x_1)^2\, d\pi_k(x_0,x_1) = W_2(\mu_{0,k}, \mu_{1,k})^2. \tag{28.43}$$

By Theorem 28.6, $(f_k)_\#$ is an approximate isometry $P_2(\mathcal{X}_k) \to P_2(\mathcal{X})$, so

$$\lim_{k\to\infty} W_2(\mu_{0,k}, \mu_{1,k})^2 = \lim_{k\to\infty} W_2\big((f_k)_\# \mu_{0,k}, (f_k)_\# \mu_{1,k}\big) = W_2(\mu_0, \mu_1)^2, \tag{28.44}$$

where the latter limit follows from the continuity of W_2 under weak convergence (Theorem 6.7). The combination of (28.42), (28.43) and (28.44) shows that $\int d(x_0,x_1)^2\, d\pi(x_0,x_1) = W_2(\mu_0,\mu_1)^2$, so π is an optimal transference plan. So $\mathit{\Pi}$ is an optimal dynamical transference plan, and the proof of (vi) is complete.

(Note: Since $(\mu_{t,k})_{0\le t\le 1}$ is a geodesic path in $P_2(\mathcal{X}_k)$ (recall Corollary 7.20), and $(f_k)_\#$ is an approximate isometry $P_2(\mathcal{X}_k) \to P_2(\mathcal{X})$, Theorem 27.9 implies directly that the limit path $\mu_t = (f_k)_\# \mu_{k,t}$ is a geodesic in $P_2(\mathcal{X})$; but I don't know if one can then recover the optimality of $\mathit{\Pi}$.) □

To complete the proof of Theorem 28.9, it only remains to establish the following lemma, which was used in the proof of statement (a).

Lemma 28.11. *Let $(\mathcal{X}, d)$ be a compact length space. Let σ be a probability measure on $[0,1] \times \mathcal{X}$ satisfying (28.28). Then there is a Lipschitz curve $\gamma : [0,1] \to \mathcal{X}$ such that $\sigma(dt\, dx) = \overline{\gamma}(dt\, dx) = \delta_{x=\gamma(t)}\, dt$.*

Proof of Lemma 28.11. First disintegrate σ with respect to its first marginal λ. Then there is a family $(\nu_t)_{0\le t\le 1}$, measurable as a map from $[0,1]$ to $P(\mathcal{X})$ and unique up to modification on a set of zero Lebesgue measure in $[0,1]$, such that

$$\sigma(dt\, dx) = \nu_t(dx)\, dt. \tag{28.45}$$

The goal is to show that, up to modification of ν_t on a negligible set of times,

$$\nu_t(dx) = \delta_{x=\gamma(t)}, \tag{28.46}$$

where γ is Lipschitz.

In the sequel I shall use, for convenience, the 1-Wasserstein distance W_1.

Step 1: almost-everywhere Lipschitz continuity

Let β be an arbitrary nonnegative continuous function on $[0,1] \times [0,1]$. Integrating (28.28) with respect to β yields

$$\int_0^1 \int_0^1 \beta(t_0, s_0)\, \mathcal{L}^\delta_{t_0 \to s_0}(\sigma)\, dt_0\, ds_0 \le C \int_0^1 \int_0^1 \beta(t_0, s_0)\, \big(|s_0 - t_0| + \delta\big)\, dt_0\, ds_0. \tag{28.47}$$

The left-hand-side of (28.47) can be written as

$$\int_0^1 \int_0^1 \int_{\mathcal{X}\times[0,1]} \int_{\mathcal{X}\times[0,1]} \beta(t_0, s_0)\, d(x,y)\, \varphi^\delta_+(t - t_0)\, \varphi^\delta_-(s - s_0)\, d\nu_t(x)\, dt\, d\nu_s(y)\, ds\, dt_0\, ds_0 \tag{28.48}$$

$$= \int_0^1 \int_0^1 F^\delta(t,s)\, \Lambda(t,s)\, dt\, ds,$$

where

$$\begin{cases} F^\delta(t,s) = \displaystyle\int_0^1 \int_0^1 \beta(t_0, s_0)\, \varphi^\delta_+(t - t_0)\, \varphi^\delta_-(s - s_0)\, dt_0\, ds_0 \\ \Lambda(t,s) = \displaystyle\int_{\mathcal{X}\times\mathcal{X}} d(x,y)\, d\nu_t(x)\, d\nu_s(y). \end{cases} \tag{28.49}$$

Since $F^\delta(t,s)$ converges to $\beta(t,s)$ in $C([0,1] \times [0,1])$ as $\delta \to 0$, the expression in (28.48) converges to

$$\int_{\mathcal{X}\times[0,1]} \int_{\mathcal{X}\times[0,1]} \beta(t,s)\, d(x,y)\, d\nu_t(x)\, dt\, d\nu_s(y)\, ds. \tag{28.50}$$

Now plug this back into (28.47) and let $\delta \to 0$ to conclude that

$$\int_{\mathcal{X}\times[0,1]} \int_{\mathcal{X}\times[0,1]} \beta(t,s)\, d(x,y)\, d\nu_t(x)\, dt\, d\nu_s(y)\, ds \le C \int_0^1 \int_0^1 \beta(t,s)\, |s - t|\, dt\, ds. \tag{28.51}$$

As β is arbitrary, we actually have

$$\int_{\mathcal{X}\times\mathcal{X}} d(x,y)\, \nu_t(dx)\, \nu_s(dy) \le C\, |t - s| \tag{28.52}$$

for $(\lambda \otimes \lambda)$-almost all (t,s) in $[0,1] \times [0,1]$. In particular,

$$W_1(\nu_t, \nu_s) \le C\, |t - s| \qquad \text{for almost all } (t,s) \in [0,1] \times [0,1]. \tag{28.53}$$

Step 2: true Lipschitz continuity

Now let us show that ν_t can be modified on a negligible set of times so that (28.53) holds for *all* $(t,s) \in [0,1] \times [0,1]$.

For small $\varepsilon > 0$ and $t \in [\varepsilon, 1 - \varepsilon]$, define

$$\nu_t^\varepsilon = \frac{1}{2\varepsilon} \int_{-\varepsilon}^{\varepsilon} \nu_{t+\tau}\, d\tau. \tag{28.54}$$

Then by Theorem 4.7,

$$W_1(\nu_t^\varepsilon, \nu_s^\varepsilon) \le \frac{1}{2\varepsilon} \int_{-\varepsilon}^{\varepsilon} W_1(\nu_{t+\tau}, \nu_{s+\tau})\, d\tau \le C|t-s| + O(\varepsilon). \tag{28.55}$$

Next, let $(\psi_k)_{k\in\mathbb{N}}$ be a countable dense subset of $C(\mathcal{X})$. For all k,

$$\int_{\mathcal{X}} \psi_k\, d\nu_t^\varepsilon = \frac{1}{2\varepsilon} \int_{t-\varepsilon}^{t+\varepsilon} \left(\int_{\mathcal{X}} \psi_k(x)\, d\nu_\tau(x) \right) d\tau. \tag{28.56}$$

Since the expression inside the parentheses is a bounded measurable function of τ, Lebesgue's density theorem ensures that as $\varepsilon \to 0$, the right-hand-side of (28.56) converges to $\int_{\mathcal{X}} \psi_k\, d\nu_t$ for almost all t. So there is a negligible subset of $[0,1]$, say N_k, such that if $t \notin N_k$ then

$$\lim_{\varepsilon\to 0} \int_{\mathcal{X}} \psi_k\, d\nu_t^\varepsilon = \int_{\mathcal{X}} \psi_k\, d\nu_t. \tag{28.57}$$

Let $N = \bigcup_{k=1}^\infty N_k$; this is a negligible subset of $[0,1]$. For all $t \notin N$, equation (28.57) holds for all k. This proves that $\lim_{\varepsilon\to 0} \nu_t^\varepsilon = \nu_t$ in the weak topology, for almost all t.

Now for arbitrary $t \in (0,1)$, there is a sequence of times $t_j \to t$, such that $\nu_{t_j}^\varepsilon$ converges to ν_{t_j} as $\varepsilon \to 0$. Then for ε and ε' sufficiently small,

$$\begin{aligned} W_1(\nu_t^\varepsilon, \nu_t^{\varepsilon'}) &\le W_1(\nu_{t_j}^\varepsilon, \nu_{t_j}^{\varepsilon'}) + 2C\,|t-t_j| \\ &\le W_1(\nu_{t_j}^\varepsilon, \nu_{t_j}) + W_1(\nu_{t_j}^{\varepsilon'}, \nu_{t_j}) + 2C\,|t-t_j|. \end{aligned} \tag{28.58}$$

It follows that $\lim_{\varepsilon,\varepsilon'\to 0} W_1(\nu_t^\varepsilon, \nu_t^{\varepsilon'}) = 0$. Since $(P(\mathcal{X}), W_1)$ is a complete metric space (Theorem 6.15), in fact $\lim_{\varepsilon\to 0} \nu_t^\varepsilon$ exists for *all* (not just almost all) $t \in (0,1)$. The limit coincides with ν_t for almost all $t \in (0,1)$, so it defines the same measure $\sigma(dt\, dx)$. Re-define ν_t on a negligible set of times if necessary, so that the limit is ν_t for all $t \in (0,1)$. Now it is possible to pass to the limit in (28.55) as $\varepsilon \to 0$, and recover $W_1(\nu_t, \nu_s) \le C|t-s|$ for all $t, s \in (0,1)$. Of course this extends to $t, s \in [0,1]$ by density.

Step 3: Conclusion

From the previous step, $W_1(\nu_t, \nu_{t_0}) \le C\delta$ if $|t - t_0| \le \delta$. It follows from the definition of $\mathcal{L}^\delta_{t_0\to s_0}$ that

$$\begin{aligned} \mathcal{L}^\delta_{t_0\to s_0}(\sigma) &= \int_{\mathcal{X}\times[0,1]} \int_{\mathcal{X}\times[0,1]} d(x,y)\, \varphi_+^\delta(t-t_0)\, \varphi_-^\delta(s-s_0)\, d\nu_{t_0}(x)\, dt\, d\nu_{s_0}(y)\, ds \;\; + O(\delta) \\ &= \int_{\mathcal{X}\times\mathcal{X}} d(x,y)\, d\nu_{t_0}(x)\, d\nu_{s_0}(y) \;\; + O(\delta). \end{aligned} \tag{28.59}$$

Plugging this back into (28.28) and taking $\delta \to 0$, we obtain

$$\int_{\mathcal{X}\times\mathcal{X}} d(x,y)\, d\nu_{t_0}(x)\, d\nu_{s_0}(y) \le C\,|s_0 - t_0|. \tag{28.60}$$

This holds for all t_0 and s_0, so we can choose $s_0 = t_0$ and obtain

$$\int_{\mathcal{X}\times\mathcal{X}} d(x,y)\, d\nu_{t_0}(x)\, d\nu_{t_0}(y) = 0. \tag{28.61}$$

This is only possible if ν_{t_0} is a Dirac measure. Hence for any $t_0 \in [0,1]$ there is a $\gamma(t_0) \in \mathcal{X}$ such that $\nu_{t_0} = \delta_{\gamma(t_0)}$. Then $d(\gamma(t), \gamma(s)) = W_1(\nu_t, \nu_s) \le C|t-s|$, so γ is Lipschitz continuous. This concludes the proof of Lemma 28.11. □

Noncompact spaces

Now it will be easy to extend the preceding results to noncompact spaces, by localization, thanks to the next lemma. Recall that $\Gamma(\mathcal{X})$ stands for the set of all (constant speed, minimizing) geodesic curves $[0,1] \to \mathcal{X}$.

Lemma 28.12. *Let $(\mathcal{X}, d)$ be a Polish length space, and let K be a compact in $\mathcal{X}$. Let Γ_K be the set of geodesic curves $[0,1] \to \mathcal{X}$ whose image lies entirely in K: in other words,*

$$\Gamma_K := \big\{\gamma \in \Gamma(\mathcal{X});\ \gamma([0,1]) \subset K\big\}.$$

Then Γ_K is itself a compact subset of $\Gamma(X)$.

Proof of Lemma 28.12. Let D be the diameter of K. The set Γ_K is the set of curves $\gamma : [0,1] \to K$ which are D-Lipschitz from $[0,1]$ to K and satisfy the equation $\mathcal{L}(\gamma) = d(\gamma(0), \gamma(1))$. By Ascoli's theorem, this set is compact for the uniform topology. □

Now comes the final result in this chapter.

Theorem 28.13. *Let $(\mathcal{X}_k, d_k, \star_k)$ be a sequence of locally compact Polish spaces converging in the pointed Gromov–Hausdorff topology to some locally compact Polish space $(\mathcal{X}, d, \star)$. Then $P_2(\mathcal{X}_k)$ converges to $P_2(\mathcal{X})$ in the geodesic local Gromov–Hausdorff topology.*

Remark 28.14. If a basepoint $\star$ is given in $\mathcal{X}$, there is a natural choice of basepoint for $P_2(\mathcal{X})$, namely $\delta_\star$. However, $P_2(\mathcal{X})$ is in general *not* locally compact, and it does not make sense to consider the *pointed* convergence of $P_2(\mathcal{X}_k)$ to $P_2(\mathcal{X})$.

Remark 28.15. Theorem 28.13 admits the following extension: If $(\mathcal{X}_k, d_k)$ converges to $(\mathcal{X}, d)$ in the geodesic local Gromov–Hausdorff topology, then also $P_2(\mathcal{X}_k)$ converges to $P_2(\mathcal{X})$ in the geodesic local Gromov–Hausdorff topology. The proof is almost the same and left to the reader.

Proof of Theorem 28.13. Let $R_\ell \to \infty$ be a given increasing sequence of positive numbers. Define

$$K^{(\ell)} = P_2\big(B_{R_\ell]}(\star)\big) \subset P_2(\mathcal{X}),$$

$$K_k^{(\ell)} = P_2\big(B_{R_\ell]}(\star_k)\big) \subset P_2(\mathcal{X}_k),$$

where the inclusion is understood in an obvious way (a probability measure on a subset of $\mathcal{X}$ can be seen as the restriction of a probability measure on $\mathcal{X}$). Since $B_{R_\ell]}(\star)$ is a compact set, $K^{(\ell)}$ is compact too, and so is $K_k^{(\ell)}$, for each k and each ℓ. Moreover, the union of all K_k is dense in $P_2(\mathcal{X})$, as a corollary of Theorem 6.15.

For each ℓ, there is a sequence $(f_k)_{k\in\mathbb{N}}$ such that each f_k is an ε_k isometry $B_{R_\ell]}(\star_k) \to B_{R_\ell]}(\star)$, where $\varepsilon_k \to 0$. From Proposition 28.7, $(f_k)_\#$ is an $\widetilde{\varepsilon}_{k,\ell}$-isometry $K_k^{(\ell)} \to K^{(\ell)}$, with

$$\widetilde{\varepsilon}_\ell \le 10\,(\varepsilon_k + \sqrt{2R_\ell \varepsilon_k}),$$

which goes to 0 as $k \to \infty$. So all the requirements of Definition 27.11 are satisfied, and $P_2(\mathcal{X}_k)$ does converge to $P_2(\mathcal{X})$ in the local Gromov–Hausdorff topology.

To check condition (iii) appearing in Definition 27.12, it is sufficient to note that any geodesic in $P_2(B_{R_\ell]}(\star_k))$ can be written as the law of a random geodesic joining points in $B_{R_\ell]}(\star_k)$, and such a geodesic is contained in $B_{2R_\ell]}(\star_k)$; so just choose ℓ' large enough that $R_{\ell'} \ge 2R_\ell$. □

Bibliographical Notes

Theorem 28.6 is taken from [247, Section 4], while Theorem 28.13 is an adaptation of [247, Appendix E]. Theorem 28.9 is new. (A part of this theorem was included in a preliminary version of [249], and later removed from that reference.)

The discussion about push-forwarding dynamical transference plans is somewhat subtle. The point of view adopted in this chapter is the following: when an approximate isometry f is given between two spaces, use it to push-forward a dynamical transference plan $\mathit{\Pi}$, via $(f\circ)_\# \mathit{\Pi}$. The advantage is that this is the same map that will push-forward the measure and the dynamical plan. The drawback is that the resulting object $(f\circ)_\# \mathit{\Pi}$ is **not** a dynamical transference plan, in fact it may not even be supported on continuous paths. This leads to the kind of technical sport that we've encountered in this chapter, embedding into probability measures on probability measures and so on.

Another option would be as follows: Given two spaces $\mathcal{X}$ and $\mathcal{Y}$, with an approximate isometry $f : \mathcal{X} \to \mathcal{Y}$, and a dynamical transference plan $\mathit{\Pi}$ on $\mathit{\Gamma}(\mathcal{X})$, define a **true** dynamical transference plan on $\mathit{\Gamma}(\mathcal{Y})$, which is a good approximation of $(f\circ)_\# \mathit{\Pi}$. The point is to construct a recipe which to any geodesic γ in $\mathcal{X}$ associates a geodesic $S(\gamma)$ in $\mathcal{Y}$ that is "close enough" to $f \circ \gamma$. This strategy was successfully implemented in the final version of [249, Appendix]; it is much simpler, and still it is quite sufficient for some purposes. The example treated in [249] is the stability of the "democratic condition" considered by Lott and myself; but certainly this simplified version will work for many other stability issues. On the other hand, I don't know if it is enough to treat such topics as the stability of general weak Ricci bounds, which will be considered in the next chapter.

The study of the kinetic energy measure and the speed field occurred to me during a parental meeting of the Crèche Le Rêve en Couleurs. My motivations for regularity estimates on the speed are explained in the bibliographic notes of Chapter 29, and come from a direction of research which I have more or less left apart for the moment. So Theorem 28.5 is still "in search of an application"; but I would be surprised if it would not prove useful some day.

29

Weak Ricci curvature bounds I: Definition and Stability

In Chapter 14 I recalled several formulations of the CD(K, N) curvature-dimension bound for a smooth manifold (M, g) equipped with a reference measure whose density (with respect to the volume element) is smooth. For instance, here is a possible formulation of CD(K, N) for $N < \infty$: For any C^2 function $\psi : M \to \mathbb{R}$, let $\mathcal{J}(t, \cdot)$ be the Jacobian determinant of $T_t : x \longmapsto \exp_x(t\nabla\psi(x))$, and let $\mathcal{D}(t, x) = \mathcal{J}(t, x)^{\frac{1}{N}}$; then, with the notation of Theorem 14.8,

$$\mathcal{D}(t, x) \geq \tau_{K,N}^{(1-t)} \mathcal{D}(0, x) + \tau_{K,N}^{(t)} \mathcal{D}(1, x). \tag{29.1}$$

How to generalize this definition in such a way that it would make sense in a possibly nonsmooth metric-measure space? This is definitely not obvious since (i) there might be no good notion of gradient, and (ii) there might be no good notion of exponential map either.

There are many definitions that one may try, but so far the only approach that yields acceptable results is the one based on displacement convexity. Recall from Chapters 16 and 17 two displacement convexity inequalities that characterize CD(K, N): Let μ_0 and μ_1 be two compactly supported (for simplification) absolutely continuous probability measures, let π be the optimal coupling of (μ_0, μ_1), let $(\rho_t\, \nu)_{0 \leq t \leq 1}$ be the displacement interpolation between $\mu_0 = \rho_0\, \nu$ and $\mu_1\, \nu$, let $(v_t)_{0 \leq t \leq 1}$ be the associated velocity field, then for all $U \in \mathcal{DC}_N$, $t \in [0, 1]$,

$$\int U(\rho_t)\, d\nu \leq (1 - t)\, U_\nu(\mu_0) + t\, U_\nu(\mu_1) - K_{N,U} \int_0^1 \rho_s(x)^{1 - \frac{1}{N}} |v_s(x)|^2\, G(s, t)\, ds, \tag{29.2}$$

$$\begin{aligned} \int U(\rho_t)\, d\nu \leq (1 - t) \int_{M \times M} U\left(\frac{\rho_0(x_0)}{\beta_{1-t}^{(K,N)}(x_0, x_1)} \right) \beta_{1-t}^{(K,N)}(x_0, x_1)\, \pi(dx_1 | x_0)\, \nu(dx_0) + \\ t \int_{M \times M} U\left(\frac{\rho_1(x_1)}{\beta_t^{(K,N)}(x_0, x_1)} \right) \beta_t^{(K,N)}(x_0, x_1)\, \pi(dx_0 | x_1)\, \nu(dx_1). \end{aligned} \tag{29.3}$$

Here $G(s, t)$ is the one-dimensional Green function of (16.6), $K_{N,U}$ is defined by (17.9), and the distortion coefficients $\beta_t^{(K,N)} = \beta_t$ are defined in (14.60).

Which of these formulas should we choose for the extension to nonsmooth spaces? When $K = 0$, both formulas reduce to just

$$\int U(\rho_t)\, d\nu \leq (1 - t) \int U(\rho_0)\, d\nu + t \int U(\rho_1)\, d\nu. \tag{29.4}$$

In the case $N < \infty$, formula (29.3) is much more convenient to establish functional inequalities; while in the case $N = \infty$ it is formula (29.2) which is easier to use. However, it will turn out that in the case $N = \infty$, (29.2) is an immediate consequence of (29.3). All this concurs to suggest that (29.3) is the correct choice on which we should base the general definition.

Now we would like to adapt these formulas to a nonsmooth context. This looks simpler than working with (29.1), but there are still a few issues to take into account.

(i) **First issue: Non-uniqueness of the displacement interpolation.**

There is a priori no reason to expect uniqueness of the displacement interpolation in a nonsmooth context. We may require the distorted displacement convexity (29.3) along every displacement interpolation, i.e. every geodesic in Wasserstein space; but this is not a good idea for stability issues. (Recall Example 27.17: in general the geodesics in the limit space cannot be realized as limits of geodesics.) Instead, we shall require a convexity inequality along *some* displacement interpolation. In this sense, we will only impose a **weak displacement convexity** property: For any μ_0 and μ_1 there should be *some* geodesic $(\mu_t)_{0\le t\le 1}$ along which inequality (29.3) should hold true.

To appreciate the difference between "convexity" and "weak convexity", note the following: If F is a function defined on a length space $\mathcal{X}$, then the two statements "F is convex along each geodesic $(\gamma_t)_{0\le t\le 1}$" and " "For any x_0 and x_1, there is a geodesic $(\gamma_t)_{0\le t\le 1}$ joining x_0 to x_1, such that $F(\gamma_t) \le (1-t)F(\gamma_0) + tF(\gamma_1)$" are not equivalent in general. (They become equivalent under some regularity assumption on $\mathcal{X}$, for instance if any two close enough points in $\mathcal{X}$ are joined by a unique geodesic.)

(ii) **Second issue: Treatment of the singular part.**

Even if μ_0 and μ_1 are absolutely continuous with respect to ν, there is no guarantee that the Wasserstein interpolant μ_t will also be absolutely continuous. For stability issues it will also be useful to work with possibly singular measures, since the set $P_2^{\rm ac}(\mathcal{X}, \nu)$ is not closed under weak convergence.

So the problem arises to devise a "correct" definition for the integral functionals of the density which appear in the displacement convexity inequalities, namely

$$U_\nu(\mu) = \int_{\mathcal{X}} U\left(\frac{d\mu}{d\nu}\right) d\nu, \qquad U^\beta_{\pi,\nu}(\mu) = \int_{\mathcal{X}\times\mathcal{X}} U\left(\frac{1}{\beta(x,y)}\frac{d\mu}{d\nu}(x)\right) \beta(x,y)\,\pi(dy|x)\,\nu(dx).$$

It would be a mistake to keep the same definition and replace $d\mu/d\nu$ by the density of the absolutely continuous part of μ with respect to ν. In fact there is only one natural extension of the functionals U and $U^\beta_{\pi,\nu}$; before giving it explicitly, I shall try to motivate it. Take a reference measure ν, a probability measure $\mu \in P_2(\mathcal{X}, \nu)$, and a convex continuous function $U : \mathbb{R}_+ \to \mathbb{R}_+$. Think of the singular part of μ as something which "always has infinite density". Assume that the respective contributions of finite and infinite values of the density decouple, so one would define separately the contributions of the absolutely continuous part $\mu_{\rm ac}$ and the singular part μ_s. Only the asymptotic behavior of $U(r)$ as $r \to \infty$ should count when one defines the contribution of μ_s. Finally, if $U(r)$ were increasing like cr, then it is natural to assume that $U_\nu(\mu_s)$ should be $\int_{\mathcal{X}} c\, d\mu_s = c\,\mu_s[\mathcal{X}]$. So it is the *asymptotic slope* of U that should matter. Since U is convex, there is a natural notion of asymptotic slope of U:

$$U'(\infty) := \lim_{r\to\infty} \frac{U(r)}{r} = \lim_{r\to\infty} U'(r) \qquad \in \mathbb{R} \cup \{+\infty\}. \tag{29.5}$$

After these preparations, the following definition should seem quite natural to the reader.

Definition 29.1 (Integral functionals for singular measures). *Let $(\mathcal{X}, d, \nu)$ be a locally compact metric-measure space, where ν is locally finite; let $U : \mathbb{R}_+ \to \mathbb{R}$ be a continuous convex function with $U(0) = 0$, and let μ be a measure on $\mathcal{X}$, compactly supported. Let*

$$\mu = \rho\,\nu + \mu_s$$

stand for the Lebesgue decomposition of μ into an absolutely continuous and a singular part. Then,

(i) define the integral functional U_ν, with nonlinearity U and reference measure ν, by

$$U_\nu(\mu) := \int_{\mathcal{X}} U(\rho(x))\,\nu(dx) \;+\; U'(\infty)\,\mu_s[\mathcal{X}];$$

(ii) if π is a probability measure on $\mathcal{X} \times \mathcal{X}$ and β is a positive measurable function on $\mathcal{X} \times \mathcal{X}$, define the integral functional $U^\beta_{\pi,\nu}$ with nonlinearity U, reference measure ν, coupling π and distortion coefficient β, by

$$U^\beta_{\pi,\nu}(\mu) := \int_{\mathcal{X}\times\mathcal{X}} U\left(\frac{\rho(x)}{\beta(x,y)}\right) \beta(x,y)\,\pi(dy|x)\,\nu(dx) \;+\; U'(\infty)\,\mu_s[\mathcal{X}],$$

where $\pi(dy|x)$ is the conditional law of y (under $\pi(dx\,dy)$) given x.

Remark 29.2. It is clear that $U^\beta_{\pi,\nu}$ reduces to U_ν when $\beta \equiv 1$, i.e. when there is no distortion.

Remark 29.3. These definitions take care of the subtleties linked to singularities of the measure μ; actually there are also subtleties linked to the behavior at infinity, but I shall consider them only in the next chapter. For the moment, we can be content with the restriction of compact support, because we know that a displacement interpolation between two compactly supported measures is itself compactly supported at all times.

For later use I record here two elementary lemmas about the functionals $U^\beta_{\pi,\nu}$. The reader may skip them at first reading and go directly to the next section.

First, there is a handy way to rewrite $U^\beta_{\pi,\nu}(\mu)$ when $\mu_s = 0$:

Lemma 29.4 (Rewriting of the distorted U_ν functional). *With the notation of Definition 29.1, if $\mu_s = 0$ then*

$$\begin{aligned} U^\beta_{\pi,\nu}(\mu) &= \int_{\mathcal{X}\times\mathcal{X}} U\left(\frac{\rho(x)}{\beta(x,y)}\right) \frac{\beta(x,y)}{\rho(x)}\,\pi(dx\,dy) \\ &= \int_{\mathcal{X}\times\mathcal{X}} v\left(\frac{\rho(x)}{\beta(x,y)}\right)\,\pi(dx\,dy), \end{aligned}$$

where $v(r) = U(r)/r$, with the convention $U(0)/0 = U'(0) \in [-\infty, +\infty)$.

Proof of Lemma 29.4. The identity is formally obvious if one notes that $\rho(x)\,\pi(dy|x)\,\nu(dx) = \pi(dy\,dx) = \pi(dy|x)\,\mu(dx)$; so all the subtlety lies in the fact that in the left-hand side, the convention is $U(0)/0 = 0$, while in the right-hand side it is $U(0)/0 = U'(0)$. Switching between both conventions is allowed because the set $\{\rho = 0\}$ is anyhow of zero π-measure. □

Secondly, the functionals $U^\beta_{\pi,\nu}$ (and the functionals U_ν) satisfy a principle of "rescaled subadditivity", which might at first sight seem contradictory with the convexity property, but is not at all. In the next statement, I use the notation

$$U_a(r) = a^{-1}\, U(ar).$$

Lemma 29.5 (Rescaled subadditivity of the distorted U_ν functionals). *Let $(\mathcal{X}, d, \nu)$ be a locally compact metric-measure space, where ν is locally finite, and let β be a positive measurable function on $\mathcal{X} \times \mathcal{X}$. Let U be a Lipschitz convex function with $U(0) = 0$. Let $\mu_1, \dots, \mu_k$ be probability measures on $\mathcal{X}$, let $\pi_1, \dots, \pi_k$ be probability measures on $\mathcal{X} \times \mathcal{X}$, and let $Z_1, \dots, Z_k$ be positive numbers with $\sum Z_j = 1$. Then*

$$U^\beta_{\sum_j Z_j\, \pi_j, \nu}\big(\sum Z_j\, \mu_j\big) \le \sum_j Z_j\, (U_{Z_j})^\beta_{\pi_j,\nu}(\mu_j),$$

with equality if the measures μ_k are singular with respect to each other.

Proof of Lemma 29.5. By induction, it is sufficient to prove the lemma when $k = 2$. We start by the following remark: If x, y are nonnegative numbers, then

$$U(x + y) \ge U(x) + U(y). \tag{29.6}$$

Inequality (29.6) follows at once from the fact that $U(t)/t$ is a nondecreasing function of t, and thus

$$\frac{U(x)}{x} \le \frac{U(x+y)}{x+y},\ \frac{U(y)}{y} \le \frac{U(x+y)}{x+y} \Longrightarrow xU(x+y) + yU(x+y) \ge (x+y)(U(x)+U(y)).$$

Now we turn to the proof of the lemma. For pedagogic reasons, I shall first treat the special case when $\beta = 1$, so that $U^\beta_{\pi,\nu} = U_\nu$. With obvious notation,

$$U_\nu(Z_1\mu_1 + Z_2\mu_2) = \int U(Z_1\rho_1 + Z_2\rho_2)\, d\nu + U'(\infty)\, \big(Z_1\, \mu_{1,s}[\mathcal{X}] + Z_2\, \mu_{2,s}[\mathcal{X}]\big);$$

$$(U_{Z_1})_\nu(\mu_1) = \frac{1}{Z_1} \int U(Z_1\rho_1)\, d\nu + U'(\infty)\, \mu_{1,s}[\mathcal{X}];$$

$$(U_{Z_2})_\nu(\mu_2) = \frac{1}{Z_2} \int U(Z_2\rho_2)\, d\nu + U'(\infty)\, \mu_{2,s}[\mathcal{X}];$$

then the conclusion follows immediately from (29.6). Moreover, the claim about equality merely amounts to say that $U(x + y) = U(x) + U(y)$ as soon as either x or y is zero.

Now for the general case, we observe that

$$\begin{aligned} U^\beta_{Z_1\pi_1 + Z_2\pi_2,\nu}&(Z_1\mu_1 + Z_2\mu_2) \\ &= \int_{\mathcal{X}\times\mathcal{X}} U\left(\frac{Z_1\rho_1(x) + Z_2\rho_2(x)}{\beta(x,y)}\right)\, \beta(x,y)(Z_1\pi_1 + Z_2\pi_2)(dy|x)\, \nu(dx) \\ &+ U'(\infty)\, \big(Z_1\, \mu_{1,s}[\mathcal{X}] + Z_2\, \mu_{2,s}[\mathcal{X}]\big); \end{aligned}$$

$$(U_{Z_1})^\beta_{\pi_1,\nu}(\mu_1) = \int U\left(\frac{Z_1\rho_1(x)}{\beta(x,y)}\right)\, \beta(x,y)\pi_1(dy|x)\, d\nu + U'(\infty)\mu_{1,s}[M];$$

$$(U_{Z_2})^\beta_{\pi_2,\nu}(\mu_2) = \int U\left(\frac{Z_2\rho_2(x)}{\beta(x,y)}\right)\, \beta(x,y)\pi_2(dy|x)\, d\nu + U'(\infty)\mu_{2,s}[M].$$

Thus the result follows again by (29.6). □

Synthetic definition of the curvature-dimension bound

In the next definition I shall use the following conventions: An optimal transference π is said to be associated with a displacement interpolation $(\mu_t)_{0\le t\le 1}$ if there is a dynamical optimal transference plan Π such that $\mu_t = (e_t)_\# \Pi$, $\pi = (e_0, e_1)_\# \pi$. (Equivalently, there is a random geodesic γ such that $\pi = \text{law}\,(\gamma_0, \gamma_1)$ and $\mu_t = \text{law}\,(\gamma_t)$.) Also, if π is a given probability measure on $\mathcal{X} \times \mathcal{X}$, I shall denote by $\check{\pi}$ the probability measure obtained from π by "exchanging x and y"; more rigorously, $\check{\pi} := S_\# \pi$, where $S(x, y) = (y, x)$.

Definition 29.6 (Weak curvature-dimension condition). *Let $K \in \mathbb{R}$ and $N \in [1, \infty]$. A locally compact, complete σ-finite metric-measure length space $(\mathcal{X}, d, \nu)$ is said to satisfy a weak* $\mathrm{CD}(K, N)$ *condition, or to be a weak* $\mathrm{CD}(K, N)$ *space, if the following condition is satisfied: Whenever ρ_0 and ρ_1 are two compactly supported probability densities, and $\mu_0 = \rho_0\, \nu$, $\mu_1 = \rho_1\, \nu$ are the associated probability measures, then there exists a displacement interpolation $(\mu_t)_{0\le t\le 1}$ and an associated optimal coupling π of (μ_0, μ_1) such that, for all $U \in \mathcal{DC}_N$ and for all $t \in [0, 1]$,*

$$U_\nu(\mu_t) \le (1-t)\, U^{\beta^{(K,N)}_{1-t}}_{\pi,\nu}(\mu_0) \;+\; t\, U^{\beta^{(K,N)}_{t}}_{\check{\pi},\nu}(\mu_1). \tag{29.7}$$

In short, the weak $\mathrm{CD}(K, N)$ condition means that the functionals U_ν should be weakly displacement convex with distortion coefficients $(\beta^{(K,N)}_t)$, for all $U \in \mathcal{DC}_N$. This is a property of the triple $(\mathcal{X}, d, \nu)$, but for simplicity I shall often abbreviate the statement "$(\mathcal{X}, d, \nu)$ satisfies a weak $\mathrm{CD}(K, N)$ condition" into "$\mathcal{X}$ satisfies a weak $\mathrm{CD}(K, N)$ condition, with the understanding that the distance and reference measure should be clear from the context.

Before going any further, I shall make explicit the fact that this definition is an extension of the usual one, and connect the synthetic notion of weak $\mathrm{CD}(K, N)$ space with the corresponding analytic notion (considered in Chapter 14, and defined for instance in terms of the modified Ricci curvature tensor (14.36)).

Theorem 29.7 (Smooth weak $\mathrm{CD}(K, N)$ **spaces are** $\mathrm{CD}(K, N)$ **manifolds).** *Let (M, g) be a smooth Riemannian manifold, equipped with its geodesic distance d, its volume measure* vol*, and a reference measure $\nu = e^{-V}$ vol, where $V \in C^2(M)$. Then, (M, d, ν) is a weak* $\mathrm{CD}(K, N)$ *space if and only if (M, g, ν) satisfies the* $\mathrm{CD}(K, N)$ *curvature-dimension bound; or equivalently, if the modified Ricci tensor* $\mathrm{Ric}_{N,\nu}$ *satisfies* $\mathrm{Ric}_{N,\nu} \ge Kg$.

Proof of Theorem 29.7. This is a direct consequence of Theorems 10.35 and 17.28. □

The end of this section is devoted to a series of comments about Definition 29.6.

• In the case $K > 0$ and $N < \infty$, the coefficient $\beta^{(K,N)}_t(x, y)$ takes the value $+\infty$ if $0 < t < 1$ and $d(x, y) > D_{K,N} := \pi\sqrt{(N-1)/K}$. In that case the natural convention in the integral is $\infty\, U(r/\infty) = U'(0)\, r$, in accordance with Lemma (29.4). With this convention, Definition 29.6 implies that *the diameter of the support of ν is automatically bounded above by $D_{K,N}$*. Otherwise, take $x_0, x_1 \in \text{Spt}\,\nu$ with $d(x_0, x_1) > D_{K,N}$, and choose $r > 0$ small enough that for all $x_0' \in B_r(x_0)$, $x_1' \in B_r(x_1)$, one has still $d(x_0', x_1') > D_{K,N}$. Take $\rho_0 = 1_{B_r(x_0)}/\nu[B_r(x_0)]$ and $\rho_1 = 1_{B_r(x_1)}/\nu[B_r(x_1)]$ in the definition of weak $\mathrm{CD}(K, N)$ bound. Then the coefficients β_t appearing in the right-hand side of (29.7) are always $+\infty$, and the measures have no singular part; so inequality (29.7) becomes just

$$U_\nu(\mu_t) \le\; U'(0)\Big((1-t)\int \rho_0\, d\nu \;+\; t\int \rho_1\, d\nu\Big) = U'(0). \tag{29.8}$$

Now choose $U(r) = -r^{1-1/N}$: Then $U'(0) = -\infty$, so inequality (29.8) implies $U_\nu(\mu) = -\infty$. On the other hand, by Jensen's inequality, $U_\nu(\mu) \geq -\nu[S]\,(\int \rho\, d\nu/\nu[S])^{1-1/N} = -\nu[S]^{1/N}$, where S stands for the support of μ; so $U_\nu(\mu)$ cannot be $-\infty$. This contradiction proves the claim. Let me record the conclusion in the form of a separate statement:

Proposition 29.8 (Bonnet–Myers diameter bound for weak $\mathrm{CD}(K,N)$ spaces). *If $(\mathcal{X}, d, \nu)$ is a weak $\mathrm{CD}(K,N)$ space with $K > 0$ and $N < \infty$, then*

$$\mathrm{diam}\,(\mathrm{Spt}\,\nu) \leq D_{K,N} := \pi\sqrt{\frac{N-1}{K}}.$$

As a corollary, when we use inequality (29.7) in a weak $\mathrm{CD}(K,N)$ space, the distortion coefficients appearing in the right-hand side are in fact always finite.

• An important property of the classical $\mathrm{CD}(K,N)$ condition is that it is more and more stringent as K increases and as N decreases. The next proposition shows that the same is true in a nonsmooth setting.

Proposition 29.9 (consistency of the $\mathrm{CD}(K,N)$ conditions). *The weak condition $\mathrm{CD}(K,N)$ is more and more stringent as K increases, and as N decreases.*

Proof. First, the class $\mathcal{DC}_N$ becomes smaller as N increases, which means less conditions to satisfy. Next, recall that $\beta_t^{(K,N)}$ and $\beta_{1-t}^{(K,N)}$ are increasing in K and decreasing in N (as noticed right after (14.60)); since $U(r)/r$ is nonincreasing, the quantities $\beta_{1-t}^{(K,N)}(x_0,x_1)U(\rho_0(x_0)/\beta_{1-t}^{(K,N)})$ and $\beta_t^{(K,N)}(x_0,x_1)U(\rho_1(x_1)/\beta_t^{(K,N)})$ are nondecreasing in N and nonincreasing in K. The conclusion follows immediately. □

• To check Definition 29.6, it is not really necessary to establish inequality (29.7) for the whole class $\mathcal{DC}_N$: It is sufficient to restrict to members of $\mathcal{DC}_N$ that are Lipschitz and nonnegative. This is the content of the next statement.

Proposition 29.10 (Sufficient condition for Definition (29.6)). *In Definition 29.6, it is equivalent to require that inequality (29.7) hold for all $U \in \mathcal{DC}_N$, or just for all $U \in \mathcal{DC}_N \cap \mathrm{Lip}(\mathbb{R}_+, \mathbb{R}_+)$.*

Proof. We shall assume that $(\mathcal{X}, d, \nu)$ satisfies Definition 29.6, except that (29.6) holds true only for $U \in \mathcal{DC}_N \cap \mathrm{Lip}(\mathbb{R}_+, \mathbb{R}_+)$, and check that in fact inequality (29.6) holds true for all $U \in \mathcal{DC}_N$. This will be done in three steps.

Step 1: Let $U \in \mathcal{DC}_N \cap \mathrm{Lip}(\mathbb{R}_+, \mathbb{R})$. Then U can be decomposed as

$$U(r) = \widetilde{U}(r) - Ar,$$

where $\widetilde{U} \in \mathcal{DC}_N \cap \mathrm{Lip}(\mathbb{R}_+, \mathbb{R}_+)$ and $A \geq 0$ (choose $A = \max(-U'(0), 0)$). By assumption, with the same notation as in Definition 29.6, one has the inequality

$$\widetilde{U}_\nu(\mu_t) \leq (1-t)\,\widetilde{U}_{\pi,\nu}^{\beta_{1-t}^{(K,N)}}(\mu_0) \;+\; t\,\widetilde{U}_{\check{\pi},\nu}^{\beta_t^{(K,N)}}(\mu_1). \tag{29.9}$$

Write $\mu_t = \rho_t\,\nu + (\mu_t)_s$ for the Lebesgue decomposition of μ_t with respect to ν. The replacement of $\widetilde{U}$ by U amounts to adding to the left-hand side

$$A\left(\int \rho_t\, d\nu + (\mu_t)_s[\mathcal{X}]\right) = A\mu_t[\mathcal{X}] = A,$$

and to the right-hand side

$$A\,(1-t)\int_{\mathcal{X}\times\mathcal{X}} \frac{\rho_0(x_0)}{\beta^{(K,N)}_{1-t}(x_0,x_1)}\,\beta^{(K,N)}_{1-t}(x_0,x_1)\,\pi(dx_1|x_0)\,\nu(dx_0)$$
$$+\,A\,t\int_{\mathcal{X}\times\mathcal{X}} \frac{\rho_1(x_1)}{\beta^{(K,N)}_{t}(x_0,x_1)}\,\beta^{(K,N)}_{t}(x_0,x_1)\,\pi(dx_0|x_1)\,\nu(dx_1)\;=A$$

also. So the inequality (29.9) also holds true with $\widetilde{U}$ replaced by U.

Step 2: Let $U\in\mathcal{DC}_N$ with $U'(\infty)<\infty$. Then by Proposition 17.7(iv), there is a nonincreasing sequence $(U_\ell)_{\ell\in\mathbb{N}}$, converging pointwise to U, with $U'_\ell(0)>-\infty$ and $U'_\ell(0)\to U'(0)$. Necessarily, $U'_\ell(\infty)<U'(\infty)<\infty$, so each U_ℓ is Lipschitz, and one has, by Step 1,

$$(U_\ell)_\nu(\mu_t)\le(1-t)\,(U_\ell)^{\beta^{(K,N)}_{1-t}}_{\pi,\nu}(\mu_0)\;+\;t\,(U_\ell)^{\beta^{(K,N)}_{t}}_{\check\pi,\nu}(\mu_1).$$

The problem is to pass to the limit as $\ell\to\infty$. In the left-hand side this is obvious, since $U\le U_\ell$. In the right-hand side, this will follow from the monotone convergence theorem as soon as we have checked that the integrands are bounded above, uniformly in ℓ, by integrable functions. But such is the case since, for instance,

$$(U_\ell)\left(\frac{\rho_0(x_0)}{\beta^{(K,N)}_{1-t}(x_0,x_1)}\right)\beta^{(K,N)}_{1-t}(x_0,x_1)\;\le(U_\ell)'(\infty)\frac{\rho_0(x_0)}{\beta^{(K,N)}_{1-t}(x_0,x_1)}\;\beta^{(K,N)}_{1-t}(x_0,x_1)$$
$$\le U'(\infty)\rho_0(x_0)\quad\in L^1(\pi(dx_0|x_1)\,\nu(dx_1)).$$

Step 3: Finally we consider the case of a general $U\in\mathcal{DC}_N$. The reasoning is pretty much the same as for Step 2. By Proposition 17.7(iii), there is a nondecreasing sequence $(U_\ell)_{\ell\in\mathbb{N}}$, converging pointwise to U, with $U'_\ell(\infty)<\infty$ and $U'_\ell(\infty)\to U'(\infty)$. By Step 2, one has

$$(U_\ell)_\nu(\mu_t)\le(1-t)(U_\ell)^{\beta^{(K,N)}_{1-t}}_{\pi,\nu}(\mu_0)\;+\;t(U_\ell)^{\beta^{(K,N)}_{1-t}}_{\check\pi,\nu}(\mu_1),$$

and it remains to pass to the limit as $\ell\to\infty$. In the right-hand side, this is obvious since $U_\ell\le U$. As for the left-hand side, it may be rewritten as

$$\int U_\ell(\rho_t)\,d\nu\;+\;U'_\ell(\infty)\,(\mu_t)_s[\mathcal{X}]. \tag{29.10}$$

Then we know that $U'_\ell(\infty)\to U'(\infty)$, so we may pass to the limit in the second term of (29.10). To pass to the limit in the first term by monotone convergence, it suffices to check that $U_\ell(\rho_t)$ is bounded below, uniformly in ℓ, by a ν-integrable function. But this is true since, for instance, U_0 which is bounded below by an affine function of the form $r\to-C(r+1)$, $C\ge0$; so $U_\ell(\rho_t)\ge-C\rho_t-C\,1_{\rho_t>0}$, and the latter function is integrable since ρ_t has compact support. □

• In Definition 29.6 I imposed μ_0 and μ_1 to be (a) compactly supported, (b) absolutely continuous with respect to ν. Both assumptions can actually be relaxed, but the definition which one gets by so doing is not stronger; see Theorem 30.8 in the next chapter for more details.

• Finally, here are some **examples** of weak $\mathrm{CD}(K,N)$ spaces.

Example 29.11. Let V be a continuous function $\mathbb{R}^n \to \mathbb{R}$ with $\int e^{-V(x)}\,dx < \infty$, and let $\nu(dx) = e^{-V(x)}\,dx$. Let d_2 be the usual Euclidean distance. Then the space $(\mathbb{R}^n, d_2, \nu)$ satisfies the usual $\mathrm{CD}(K,\infty)$ condition if V is C^2 and $\nabla^2 V \geq K\,I_n$ in the classical sense. It satisfies the weak $\mathrm{CD}(K,\infty)$ condition *without any regularity assumption on* V, as soon as $\nabla^2 V \geq K\,I_n$ in the sense of distributions, which means that V is K-convex. For instance, if V is merely convex, then $(\mathbb{R}^n, d_2, \nu)$ satisfies the weak $\mathrm{CD}(0,\infty)$ condition. To see this, note that if $\mu(dx) = \rho(x)\,dx$, then

$$H_\nu(\mu) = \int \rho(x) \log \rho(x)\,dx + \int \rho(x)\,V(x)\,dx = H(\mu) + \int V\,d\mu;$$

then the first term is always displacement convex, and the second is displacement convex if V is convex (simple exercise).

On the contrary, if V is not convex, then one can find $x_0, x_1 \in \mathbb{R}^n$ and $t \in [0,1]$ such that

$$V\big((1-t)x_0 + t x_1\big) > (1-t)\,V(x_0) + t\,V(x_1).$$

Now let ρ be a compactly supported probability density, and $\rho^\epsilon = \epsilon^{-n}\rho(\cdot/\epsilon)$: then $\rho^\epsilon(x)\,dx$ converges weakly to δ_0, so for ϵ small enough,

$$\int V(x)\,\rho^\epsilon\big(x - (1-t)x_0 - t x_1\big)\,dx > (1-t)\int V(x)\,\rho^\epsilon(x-x_0)\,dx + t\int V(x)\,\rho^\epsilon(x-x_1)\,dx.$$

On the other hand, $\int \rho^\epsilon(x-v)\log\rho^\epsilon(x-v)\,dx$ is independent of $v \in \mathbb{R}^n$; so

$$H_{e^{-V}\,dx}\big(\rho^\epsilon(\cdot - (1-t)x_0 - t x_1)\big)\,dx \;>\; (1-t)\,H_{e^{-V}\,dx}\big(\rho^\epsilon(\cdot - x_0)\big) \;+\; t\,H_{e^{-V}\,dx}\big(\rho^\epsilon(\cdot - x_1)\big).$$

Since the path $(\rho^\epsilon(x-(1-s)x_0 - s x_1)\,dx)_{0\leq s\leq 1}$ is a geodesic interpolation (this is the translation at uniform speed, corresponding to $\nabla\psi = \text{constant}$), we see that $(\mathbb{R}^n, d_2, e^{-V(x)}\,dx)$ cannot be a weak $\mathrm{CD}(0,\infty)$ space. The conclusion is that $(\mathbb{R}^n, d_2, e^{-V(x)}\,dx)$ satisfies a weak $\mathrm{CD}(0,\infty)$ condition *if and only if* V is convex.

Example 29.12. Let M be a smooth n-dimensional compact Riemannian manifold with nonnegative Ricci curvature, and let G be a compact Lie group acting isometrically on M. (See the bibliographical notes for references on these notions.) Let $\mathcal{X} = M/G$ and let $q : M \to \mathcal{X}$ be the quotient map. Equip $\mathcal{X}$ with the distance $d(x,y) = \inf\{d_M(x',y');\ q(x') = x, q(y') = y\}$, and with the measure $\nu = q_\# \mathrm{vol}_M$. Then the resulting space $(\mathcal{X}, d, \nu)$ is a weak $\mathrm{CD}(0,n)$ space, that in general will not be a manifold. (There will typically be singularities at fixed points of the group action.)

Example 29.13. It will be shown later on that $(\mathbb{R}^n, \|\cdot\|, \lambda_n)$ is a weak $\mathrm{CD}(K,N)$ space, where $\|\cdot\|$ is *any* norm on $\mathbb{R}^n$, and λ_n is the n-dimensional Lebesgue measure. This example proves that a weak $\mathrm{CD}(K,N)$ space may be "strongly" branching (recall the discussion in Example 27.17).

The remaining part of this chapter is devoted to a proof of stability for the weak $\mathrm{CD}(K,N)$ property.

Some properties of the integral functionals U_ν and $U^\beta_{\pi,\nu}$

In this section, I shall explain some of the remarkable properties of the integral functionals appearing in Definition 29.1. For the moment it will be sufficient to restrict to the case

of a *compact* space $\mathcal{X}$, and it will be convenient to consider that U_ν and $U^\beta_{\pi,\nu}$ are defined on the set of all (nonnegative) finite Borel measures, not necessarily *probability* measures. (Actually, in Definition 29.1 it was not assumed that μ is a probability measure.) One may even think of these functionals as defined on the whole vector space $M(\mathcal{X})$ of finite Borel measures on $\mathcal{X}$, with the convention that their value is $+\infty$ if μ is not nonnegative; then U_ν and $U^\beta_{\pi,\nu}$ are true convex functional on $M(\mathcal{X})$.

It is convenient to study the functionals U_ν by means of their **Legendre representation**. Generally speaking, the Legendre representation of a convex functional Φ defined on a vector space E is an identity of the form

$$\Phi(x) = \sup\Big\{ \langle \Lambda, x\rangle - \Psi(\Lambda) \Big\},$$

where Λ varies over a certain subset of E^*, and Ψ is a convex functional of Λ. Usually, Λ varies over the whole set E^*, and $\Psi(\Lambda) = \sup_{x\in E}[\langle \Lambda, x\rangle - \Phi(x)]$ is the Legendre transform of Φ; but here we don't really want to do so, because nobody knows what the huge space $M(\mathcal{X})^*$ looks like. So it is better to restrict to subspaces of $M(\mathcal{X})^*$. There are several natural possible choices, resulting in various Legendre representations; which one is most convenient depends on the context. Here below are the ones that will be useful in the sequel.

Definition 29.14 (Legendre transform of a real-valued convex function). *Let $U : \mathbb{R}_+ \to \mathbb{R}$ be a continuous convex function with $U(0) = 0$; its Legendre transform is defined on $\mathbb{R}$ by*

$$U^*(p) = \inf_{r\in\mathbb{R}_+} \Big[pr - U(r) \Big].$$

This is a convex function, taking the value $-U(0) = 0$ on $(-\infty, U'(0))$ and $+\infty$ on $(U'(\infty), +\infty)$.

Proposition 29.15 (Legendre representation of U_ν). *Let $U : \mathbb{R}_+ \to \mathbb{R}$ be a continuous convex function with $U(0) = 0$, let $\mathcal{X}$ be a compact metric space, equipped with a finite reference measure ν. Then, whenever μ is a finite measure on $\mathcal{X}$,*

$$(i) \qquad U_\nu(\mu) = \sup \left\{ \int_{\mathcal{X}} \varphi \, d\mu - \int_{\mathcal{X}} U^*(\varphi)\, d\nu; \quad \varphi \in L^\infty(\mathcal{X}); \quad \varphi \le U'(\infty) \right\}$$

$$(ii) \qquad U_\nu(\mu) = \sup \left\{ \int_{\mathcal{X}} \varphi \, d\mu - \int_{\mathcal{X}} U^*(\varphi)\, d\nu; \quad \varphi \in C(\mathcal{X}), \quad 0 < \varphi < U'(\infty) \right\}$$

The deceiving simplicity of these formulas hides some subtleties: For instance, it is in general impossible to drop the restriction $\varphi \le U'(\infty)$ in (i), so the supremum is not taken over the whole vector space $L^\infty(\mathcal{X})$ but only on a subspace thereof. Proposition 29.15 can be proven by elementary tools of measure theory; see the bibliographical notes for references and comments.

In the next statement, $M_+(\mathcal{X})$ will stand for the set of finite (nonnegative) Borel measures on $\mathcal{X}$, and $L^1_+(\nu)$ for the set of nonnegative ν-integrable measurable functions on $\mathcal{X}$.

Theorem 29.16 (lower semi-continuity and contraction property of U_ν). *Let $(\mathcal{X}, d)$ be a compact metric space, equipped with a finite measure ν. Let $U : \mathbb{R}_+ \to \mathbb{R}_+$ be*

a convex continuous function, with $U(0) = 0$. Let further $\beta(x, y)$ be a continuous positive function on $\mathcal{X} \times \mathcal{X}$. Then, with the notation of Definition 29.1,

(i) $U_\nu(\mu)$ is a weakly lower semi-continuous function of both μ and ν in $M_+(\mathcal{X})$. More explicitly, if $\mu_k \to \mu$ and $\nu_k \to \nu$ in the weak topology of convergence against bounded continuous functions, then

$$U_\nu(\mu) \le \liminf_{k\to\infty} U_{\nu_k}(\mu_k).$$

(ii) U_ν satisfies a contraction principle in both μ and ν; that is, if $\mathcal{Y}$ is another compact space, and $f : \mathcal{X} \to \mathcal{Y}$ is any measurable function, then

$$U_{f_\#\nu}(f_\#\mu) \le U_\nu(\mu).$$

(iii) If U is Lipschitz continuous on $\mathbb{R}_+$, then $U^\beta_{\pi,\nu}(\mu)$ is jointly continuous as a function of (π, μ) in $P(\mathcal{X} \times \mathcal{X}) \times L^1_+(\nu)$. More explicitly, if π_k converges to π in $P(\mathcal{X} \times \mathcal{X})$ equipped with the weak topology of convergence against bounded continuous functions, and f_k converges to f in the $L^1(\nu)$ norm, then

$$U^\beta_{\pi_k,\nu}(f_k\,\nu) \longrightarrow [k \to \infty] U^\beta_{\pi,\nu}(f\,\nu).$$

Proof of Theorem 29.16. To prove (i), note that U^* is continuous on $(U'(0), U'(\infty))$; so if φ is continuous with values in $(U'(0), U'(\infty))$, then $U^*(\varphi)$ is also continuous. So Proposition 29.15(ii) can be rewritten as

$$U_\nu(\mu) = \sup_{(\varphi,\psi)\in\mathcal{U}} \left\{ \int \varphi\, d\mu + \int \psi\, d\nu \right\},$$

where $\mathcal{U}$ varies in a certain subset of $C(\mathcal{X}) \times C(\mathcal{X})$. In particular, $U_\nu(\mu)$ is a supremum of weakly continuous functions of (μ, ν); it follows that U_ν is lower semi-continuous.

To prove (ii), pick up any $\varphi \in L^\infty(\mathcal{X})$ with $\varphi \le U'(\infty)$. Then

$$\int_{\mathcal{X}} (\varphi \circ f)\, d\mu - \int_{\mathcal{X}} U^*(\varphi \circ f)\, d\nu = \int_{\mathcal{Y}} \varphi\, d(f_\#\mu) - \int_{\mathcal{Y}} U^*(\varphi)\, d(f_\#\nu).$$

If $\varphi \le U'(\infty)$, then also $\varphi \circ f \le U'(\infty)$; similarly, if φ is bounded, then also $\varphi \circ f$ is bounded. So

$$\sup_{\psi\in L^\infty;\ \varphi\le U'(\infty)} \left\{ \int_{\mathcal{X}} \psi\, d\mu - \int_{\mathcal{X}} U^*(\psi)\, d\nu \right\} \le \sup_{\varphi\in L^\infty;\ \varphi\le U'(\infty)} \left\{ \int_{\mathcal{Y}} \varphi\, d(f_\#\mu) - \int_{\mathcal{Y}} U^*(\varphi)\, d(f_\#\nu) \right\}.$$

By Proposition 29.15(i), the left-hand side coincides with $U_\nu(\mu)$, and the right-hand side with $U_{f_\#\mu}(f_\#\nu)$. This concludes the proof of the contraction property (ii).

It remains to prove (iii). If $\beta \equiv 1$ (no distortion), this is immediate: $U^\beta_{\pi,\nu}$ does not depend on π, and if $\rho, \widetilde{\rho}$ are any two functions in $L^1_+(\nu)$, one can write

$$|U_\nu(\rho) - U_\nu(\widetilde{\rho})| = \left| \int [U(\rho) - U(\widetilde{\rho})]\, d\nu \right| \le \int |U(\rho) - U(\widetilde{\rho})|\, d\nu \le \|U\|_{\mathrm{Lip}} \int |\rho - \widetilde{\rho}|\, d\nu.$$

This shows that the Lipschitz norm of U_ν on $L^1_+(\mathcal{X}, \nu)$ is bounded by the Lipschitz norm of U.

In the general case where β is variable, we can still write

$$\begin{aligned}\left|U^\beta_{\pi,\nu}(\rho\,\nu) - U^\beta_{\pi,\nu}(\widetilde{\rho}\,\nu)\right| &= \left|\int_{\mathcal{X}\times\mathcal{X}} \beta(x,y)\,\left[U\left(\frac{\rho(x)}{\beta(x,y)}\right) - U\left(\frac{\widetilde{\rho}(x)}{\beta(x,y)}\right)\right]\,\pi(dy|x)\,\nu(dx)\right| \\ &\le \|U\|_{\mathrm{Lip}} \int |\rho(x) - \widetilde{\rho}(x)|\,\pi(dy|x)\,\nu(dx) \\ &= \|U\|_{\mathrm{Lip}}\,\|\rho - \widetilde{\rho}\|_{L^1(\nu)};\end{aligned}$$

this shows that $U^\beta_{\pi,\nu}$ is still $\|U\|_{\mathrm{Lip}}$-Lipschitz in $L^1(\nu)$ for *fixed* π.

Next, if ρ is *continuous*, then according to Lemma 29.4,

$$U^\beta_{\pi,\nu}(\rho\,\nu) = \int v\left(\frac{\rho(x)}{\beta(x,y)}\right)\,\pi(dx\,dy) = \int h(x,y)\,\pi(dx\,dy),$$

where h is continuous as the composition of continuous functions. So, by definition of the weak topology, $U^\beta_{\pi,\nu}(\rho\,\nu)$ is a continuous function of π.

We can now prove statement (iii). Let ρ_k be a sequence of nonnegative L^1 functions converging to ρ in L^1, and let π_k be a sequence of probability measures on $\mathcal{X}\times\mathcal{X}$ converging weakly to π. Let $\delta > 0$ be arbitrarily small, and let $\overline{\rho}$ be a continuous density such that $\|\rho - \overline{\rho}\|_{L^1(\nu)} \le \delta$. Then

$$\begin{aligned}|U^\beta_{\pi_k,\nu}(\rho_k\,\nu) - U^\beta_{\pi,\nu}(\rho\,\nu)| \le\ & |U^\beta_{\pi_k,\nu}(\rho_k\,\nu) - U^\beta_{\pi_k,\nu}(\rho\,\nu)| + |U^\beta_{\pi_k,\nu}(\rho\,\nu) - U^\beta_{\pi_k,\nu}(\overline{\rho}\,\nu)| \\ &+ |U^\beta_{\pi_k,\nu}(\overline{\rho}\,\nu) - U^\beta_{\pi,\nu}(\overline{\rho}\,\nu)| + |U^\beta_{\pi,\nu}(\overline{\rho}\,\nu) - U^\beta_{\pi,\nu}(\rho\,\nu)|.\end{aligned}$$

The first term on the right is bounded by $\|U\|_{\mathrm{Lip}}\|\rho - \rho_k\|_{L^1}$, the second one by $\|U\|_{\mathrm{Lip}}\|\rho - \overline{\rho}\|_{L^1}$, the third one goes to 0 as $k \to \infty$, and the fourth one is bounded by $\|U\|_{\mathrm{Lip}}\|\rho - \overline{\rho}\|_{L^1}$ again, so

$$\limsup_{k\to\infty} \left|U^\beta_{\pi_k,\nu}(\rho_k\,\nu) - U^\beta_{\pi,\nu}(\rho\,\nu)\right| \le 2\|U\|_{\mathrm{Lip}}\|\rho - \overline{\rho}\|_{L^1(\nu)} \le 2\|U\|_{\mathrm{Lip}}\,\delta.$$

Since δ is arbitrarily small, the proof of (iii) is complete. □

Stability of Ricci bounds

Now we have all the tools to prove the main result in this chapter: *The weak curvature-dimension bound* $\mathrm{CD}(K,N)$ *passes to the limit.* Once again, the compact case will imply the general statement.

Theorem 29.17 (Stability of $\mathrm{CD}(K,N)$ under MGH). *Let $(\mathcal{X}_k, d_k, \nu_k)_{k\in\mathbb{N}}$ be a sequence of compact metric-measure length spaces converging in the measured Gromov–Hausdorff topology to a compact metric-measure space $(\mathcal{X}, d, \nu)$. Let $K \in \mathbb{R}$ and $N \in [1,\infty]$. If each $(\mathcal{X}_k, d_k, \nu_k)$ satisfies the weak curvature-dimension condition $\mathrm{CD}(K,N)$, then also $(\mathcal{X}, d, \nu)$ satisfies $\mathrm{CD}(K,N)$.*

Theorem 29.18. *[Stability of $\mathrm{CD}(K,N)$ under pMGH] Let $(\mathcal{X}_k, d_k, \nu_k, \star_k)$ be a sequence of locally compact, complete, separable σ-finite pointed metric-measure length spaces converging in the pointed measured Gromov–Hausdorff topology to a locally compact, complete separable σ-finite metric-measure space $(\mathcal{X}, d, \nu, \star)$. Let $K \in \mathbb{R}$ and $N \in [1,\infty]$. If each $(\mathcal{X}_k, d_k, \nu_k)$ satisfies $\mathrm{CD}(K,N)$, then also $(\mathcal{X}, d, \nu)$ satisfies $\mathrm{CD}(K,N)$.*

Remark 29.19. There is an easy generalization of Theorem 29.18: If $(\mathcal{X}_k, d_k, \nu_k)$ converges to $(\mathcal{X}, d, \nu)$ in the geodesic local Gromov–Hausdorff topology, and each $(\mathcal{X}_k, d_k, \nu_k)$ satisfies $\mathrm{CD}(K,N)$, then also $(\mathcal{X}, d, \nu)$ satisfies $\mathrm{CD}(K,N)$.

Proof of Theorem 29.17. Let $(\mathcal{X}_k, d_k, \nu_k)_{k\in\mathbb{N}}$ be a sequence of metric-measure spaces satisfying the assumptions of Theorem 29.17.

From the characterization of measured Gromov–Hausdorff convergence, we know that there are measurable functions $f_k : \mathcal{X}_k \to \mathcal{X}$ such that

(i) f_k is a ε_k-isometry $(\mathcal{X}_k, d_k) \to (\mathcal{X}, d)$, with $\varepsilon_k \to 0$;

(ii) $(f_k)_\# \nu_k$ converges weakly to ν.

Let ρ_0, ρ_1 be two probability densities on $(\mathcal{X}, \nu)$; let $\mu_0 = \rho_0\, \nu$, $\mu_1 = \rho_1\, \nu$.

For $t_0 \in \{0, 1\}$, the probability measure ρ_{t_0} can be approximated in $L^1(\nu)$ by *continuous* probability densities. (A proof of this classical fact is recalled in the bibliographical notes.) So there is a family of continuous probability densities $(\rho_{\varepsilon,t_0})_{\varepsilon\in(0,1)}$ approaching ρ_{t_0} as $\varepsilon \to 0$.

Then, still for $t_0 \in \{0, 1\}$, define

$$\mu^k_{\varepsilon,t_0} := \frac{(\rho_{\varepsilon,t_0} \circ f_k)\, \nu_k}{Z_k}, \qquad Z_k = \int (\rho_{\varepsilon,t_0} \circ f_k)\, d\nu_k.$$

Since ρ_{ε,t_0} is continuous and $(f_k)_\# \nu_k$ converges weakly to ν, we have

$$Z_k = \int \rho_{\varepsilon,t_0}\, d((f_k)_\# \nu_k) \xrightarrow[k\to\infty]{} \int \rho_{\varepsilon,t_0}\, d\nu = 1;$$

in particular $Z_k > 0$ for k large enough, and then μ^k_{ε,t_0} is a probability measure on $\mathcal{X}_k$.

Let $\psi \in C(\mathcal{X})$, then

$$\begin{aligned}\int \psi\, d((f_k)_\# \mu^k_{\varepsilon,t_0}) = \int (\psi \circ f_k)\, d\mu^k_{\varepsilon,t_0} &= \frac{1}{Z_k} \int (\psi \circ f_k)\, (\rho_{\varepsilon,t_0} \circ f_k)\, d\nu_k \\ &= \frac{1}{Z_k} \int \psi \rho_{\varepsilon,t_0}\, d((f_k)_\# \nu_k). \end{aligned} \tag{29.11}$$

On one hand, Z_k converges to 1; on the other hand

$$\int \psi \rho_{\varepsilon,t_0}\, d((f_k)_\# \nu_k) \xrightarrow[k\to\infty]{} \int \psi \rho_{\varepsilon,t_0}\, d\nu = \int \psi\, d\mu_{\varepsilon,t_0}.$$

Plugging this information back in (29.11), we obtain

$$(f_k)_\# \mu^k_{\varepsilon,t_0} \xrightarrow[k\to\infty]{} \mu_{\varepsilon,t_0} \qquad \text{weakly.} \tag{29.12}$$

Since each $(\mathcal{X}_k, d_k, \nu_k)$ satisfies $\mathrm{CD}(K, \infty)$, there is a Wasserstein geodesic $(\mu^k_{\varepsilon,t})_{0\le t\le 1}$, joining $\mu^k_{\varepsilon,0}$ to $\mu^k_{\varepsilon,1}$, such that, for all $U \in \mathcal{DC}_N$ and $t \in (0, 1)$,

$$U_{\nu_k}(\mu^k_{\varepsilon,t}) \le (1-t) U^{\beta^{(K,N)}_{1-t}}_{\pi^k_\varepsilon, \nu_k}(\mu^k_0) \;+\; t U^{\beta^{(K,N)}_t}_{\check{\pi}^k_\varepsilon, \nu}(\mu^k_1), \tag{29.13}$$

where $\beta^{(K,N)}_t$ is given by (14.60) (with the distance d_k), and π^k_ε is an optimal coupling between $\mu^k_{\varepsilon,0}$ and $\mu^k_{\varepsilon,1}$.

By Theorem 7.20, for each $\varepsilon \in (0, 1)$ and $k \in \mathbb{N}$ there is a dynamical optimal transference plan Π^k_ε such that

$$\mu^k_{\varepsilon,t} = (e_t)_\# \Pi^k_\varepsilon, \qquad \pi^k_\varepsilon = (e_0, e_1)_\# \Pi^k_\varepsilon,$$

where e_t is the evaluation at time t.

By Theorem 28.9, up to extraction of a subsequence in k, there is a dynamical optimal transference plan Π_ε on $\Gamma(\mathcal{X})$ such that, as $k \to \infty$,

$$\begin{cases} (f_k \circ)_\# \Pi_\varepsilon^k \longrightarrow \Pi_\varepsilon \qquad \text{weakly in } P(P([0,1] \times \mathcal{X})); \\ (f_k, f_k)_\# \pi_\varepsilon^k \longrightarrow \pi_\varepsilon \qquad \text{weakly in } P(\mathcal{X} \times \mathcal{X}); \\ \sup_{0 \le t \le 1} W_2\big((f_k)_\# \mu_{\varepsilon,t}^k, \mu_{\varepsilon,t}\big) \longrightarrow 0; \end{cases}$$

where

$$\mu_{\varepsilon,t} = (e_t)_\# \Pi_\varepsilon, \qquad \pi_\varepsilon = (e_0, e_1)_\# \Pi_\varepsilon.$$

Each curve $(\mu_{\varepsilon,t})_{0 \le t \le 1}$ is D-Lipschitz, where D is the diameter of $\mathcal{X}$; so by Ascoli's theorem, from $\varepsilon \in (0,1)$ we may extract a subsequence (still denoted ε for simplicity) such that

$$\sup_{0 \le t \le 1} W_2\big(\mu_{\varepsilon,t}, \mu_t\big) \xrightarrow[\varepsilon \to 0]{} 0, \tag{29.14}$$

where $(\mu_t)_{0 \le t \le 1}$ is a Wasserstein geodesic joining μ_0 to μ_1. It remains to "pass to the limit" in inequality (29.13), letting first $k \to \infty$, then $\varepsilon \to 0$, in order to show that

$$U_\nu(\mu_t) \le (1-t)\, U_{\pi,\nu}^{\beta_{1-t}^{(K,N)}}(\mu_0) \;+\; t\, U_{\check{\pi},\nu}^{\beta_t^{(K,N)}}(\mu_1). \tag{29.15}$$

By Proposition 29.10, it is sufficient to do so for $U \in \mathcal{DC}_N \cap \mathrm{Lip}(\mathbb{R}_+, \mathbb{R}_+)$. So in the sequel, t will be an arbitrary time in $(0,1)$ and U will be an arbitray nonlinearity in $\mathcal{DC}_N \cap \mathrm{Lip}(\mathbb{R}_+, \mathbb{R}_+)$.

By the joint lower semi-continuity of $(\mu, \nu) \longmapsto U_\nu(\mu)$ (Theorem 29.16(i)) and the contraction property (Theorem 29.16(ii)), we have

$$\begin{aligned} U_\nu(\mu_{\varepsilon,t}) &\le \liminf_{k \to \infty} U_{(f_k)_\# \nu_k}\big((f_k)_\# \mu_{\varepsilon,t}^k\big) \\ &\le \liminf_{k \to \infty} U_{\nu_k}(\mu_{\varepsilon,t}^k). \end{aligned} \tag{29.16}$$

Then by lower semi-continuity again,

$$U_\nu(\mu_t) \le \liminf_{\varepsilon \to 0} U_\nu(\mu_{\varepsilon,t}). \tag{29.17}$$

Inequalities (29.16) and (29.17) take care of the left-hand side of (29.13). It remains to pass to the limit in the right-hand side.

Let us consider for instance the first term in the right-hand side of (29.13), namely

$$\int \beta(x_0, y_0)\, U\left(\frac{\rho_{\varepsilon,0}^k(x_0)}{\beta(x_0, x_1)}\right) \pi_\varepsilon^k(dx_1|x_0)\, \nu(dx_0),$$

where $\beta(x,y) := \beta_{1-t}^{(K,N)}(x,y)$. The problem is to show that

$$\begin{aligned} \limsup_{\varepsilon \to 0} \limsup_{k \to \infty} \int_{\mathcal{X}_k \times \mathcal{X}_k} \beta(x_0, y_0)\, U\left(\frac{\rho_{\varepsilon,0}^k(x_0)}{\beta(x_0, x_1)}\right) \pi_\varepsilon^k(dx_1|x_0)\, \nu(dx_0) \\ \le \int_{\mathcal{X} \times \mathcal{X}} \beta(x_0, y_0)\, U\left(\frac{\rho_0(x_0)}{\beta(x_0, x_1)}\right) \pi(dx_1|x_0)\, \nu(dx_0). \end{aligned} \tag{29.18}$$

For a start, assume that $N > 1$ and $K \le 0$, so that $\beta(x,y) := \beta_{1-t}^{(K,N)}(x,y)$ is a smooth function, bounded from above and below by positive constants.

According to Lemma 29.4, the left-hand side of (29.18) can be rewritten as

$$\int_{\mathcal{X}_k} v\left(\frac{\rho^k_{\varepsilon,0}(x_0)}{\beta(x_0,x_1)}\right)\,\pi^k_\varepsilon(dx_0\,dx_1),$$

where $v(r) = U(r)/r$, extended by continuity at $r = 0$.

Since $\beta(x,y)$ is a continuous function of the distance $d_k(x,y)$, since

$$\lim_{k\to\infty}\sup_{x,y\in\mathcal{X}_k}\big|d_k(x,y) - d(f_k(x),f_k(y))\big| = 0,$$

and since $\rho^k_{\varepsilon,0}$ and v are continuous, we know that the functions $v(\rho^k_{\varepsilon,0}(x_0)/\beta(x_0,x_1))$ and $v(\rho^k_{\varepsilon,0}(x_0)/\beta(f_k(x_0),f_k(x_1)))$ are uniformly close to each other as $k\to\infty$. So

$$\lim_{k\to 0}\left|\int_{\mathcal{X}_k} v\left(\frac{\rho^k_{\varepsilon,0}(x_0)}{\beta(x_0,x_1)}\right)\,\pi^k_\varepsilon(dx_0\,dx_1) - \int_{\mathcal{X}} v\left(\frac{\rho^k_{\varepsilon,0}(x_0)}{\beta(f_k(x_0),f_k(x_1))}\right)\,\pi^k_\varepsilon(dx_0\,dx_1)\right| = 0. \quad (29.19)$$

(Of course, in the second integral the coefficient β is computed with the distance d, while in the first integral it is computed with the distance d_k.)

Then

$$\begin{aligned}\int v\left(\frac{\rho^k_{\varepsilon,0}(x_0)}{\beta(f_k(x_0),f_k(x_1))}\right)\,\pi^k_\varepsilon(dx_0\,dx_1) &= \int v\left(\frac{\rho_{\varepsilon,0}(f_k(x_0))}{\beta\big(f_k(x_0),f_k(x_1)\big)}\right)\,\pi^k_\varepsilon(dx_0\,dx_1)\\ &= \int v\left(\frac{\rho_{\varepsilon,0}(y_0)}{\beta(y_0,y_1)}\right)\,d\big[(f_k,f_k)_\#\pi^k_\varepsilon\big](y_0,y_1).\end{aligned}$$

Since $(f_k,f_k)_\#\pi^k_\varepsilon$ converges weakly to π_ε as $k\to\infty$, and since $v(\rho_{\varepsilon,0}/\beta)$ is continuous,

$$\int_{\mathcal{X}\times\mathcal{X}} v\left(\frac{\rho_{\varepsilon,0}(y_0)}{\beta(y_0,y_1)}\right)\,d\big[(f_k,f_k)_\#\pi^k_\varepsilon\big](y_0,y_1) \;\xrightarrow[k\to\infty]{}\; \int_{\mathcal{X}\times\mathcal{X}} v\left(\frac{\rho_{\varepsilon,0}(y_0)}{\beta(y_0,y_1)}\right)\,\pi_\varepsilon(dy_0\,dy_1).$$

At this stage, the conclusion is that

$$\int_{\mathcal{X}_k\times\mathcal{X}_k}\beta(x_0,y_0)\,U\left(\frac{\rho^k_{\varepsilon,0}(x_0)}{\beta(x_0,x_1)}\right)\,\pi^k_\varepsilon(dx_1|x_0)\,\nu(dx_0) \xrightarrow[k\to\infty]{} \int_{\mathcal{X}\times\mathcal{X}}\beta(x_0,y_0)\,U\left(\frac{\rho_{\varepsilon,0}(x_0)}{\beta(x_0,x_1)}\right)\,\pi_\varepsilon(dx_1|x_0)\,\nu(dx_0).$$

To finish the proof of (29.18), it is sufficient to show that

$$\int_{\mathcal{X}\times\mathcal{X}}\beta(x_0,y_0)\,U\left(\frac{\rho_{\varepsilon,0}(x_0)}{\beta(x_0,x_1)}\right)\,\pi_\varepsilon(dx_1|x_0)\,\nu(dx_0) \xrightarrow[\varepsilon\to 0]{} \int_{\mathcal{X}\times\mathcal{X}}\beta(x_0,y_0)\,U\left(\frac{\rho_0(x_0)}{\beta(x_0,x_1)}\right)\,\pi(dx_1|x_0)\,\nu(dx_0).$$

But this follows directly from Theorem 29.16(iii).

This concludes the proof of the theorem in the case $K \le 0$. Now in the case $K > 0$, a new difficulty has to be overcome: The function $\beta_t^{(K,N)}(x,y)$ diverges when the distance between x and y approaches the limit Bonnet–Myers diameter allowed by the $\mathrm{CD}(K,N)$ condition, namely $D_{K,N} := \pi\sqrt{(N-1)/K}$. Of course the problem does not arise if the supremum of the diameters of all spaces $(\mathcal{X}_k, d_k)$ is strictly less than $D_{K,N}$. Since $\mathrm{CD}(K,N)$ implies $\mathrm{CD}(K',N)$ for any $K' < K$, and then $D_{K,N} < D_{K',N}$, at least we can conclude that $(\mathcal{X}, d, \nu)$ satisfies $\mathrm{CD}(K',N)$ for any $K' < K$. Then it remains to pass to the limit in the inequality for $\mathrm{CD}(K',N)$ as $K' \uparrow K$; this can be done by monotone convergence, since e.g. $\beta_{1-t}^{(K,N)}(x_0,x_1)\, U\big(\rho_0(x_0)/\beta_{1-t}^{(K,N)}(x_0,x_1)\big)$ is a monotone function of K, as noticed in the proof of Proposition 29.9. We conclude as before that $(\mathcal{X}, d, \nu)$ satisfies $\mathrm{CD}(K,N)$.

Also the case $N = 1$ can be treated by a similar limiting argument: if all spaces $(\mathcal{X}_k, d_k, \nu_k)$ satisfy $\mathrm{CD}(K,1)$, then the limit space $(\mathcal{X}, d, \nu)$ will satisfy $\mathrm{CD}(K,N')$ for all $N' > 1$ and the conclusion will follow by letting N' decrease to 1. □

Remark 29.20. What the proof above really shows is that under certain assumptions the property of distorted displacement convexity is stable under measured Gromov–Hausdorff convergence. The usual displacement convexity is a particular case of this fact (take the distortion coefficients identically equal to 1).

Proof of Theorem 29.18. The same arguments as in the proof of Theorem 29.17 will work here, provided that we can localize the problem. So pick up probability densities ρ_0 and ρ_1 with compact support, and define $\mu_{\varepsilon,t}^k$ in exactly the same way as in the proof of Theorem 29.17. Let R be such that the supports of ρ_0 and ρ_1 are included in $B_{R]}(\star)$; then for k large enough and ε small enough, the supports of $\mu_{\varepsilon,0}^k$ and $\mu_{\varepsilon,1}^k$ are contained in $B_{R+1]}(\star_k)$. So a geodesic which starts from the support of $\mu_{\varepsilon,0}^k$ and ends in the support of $\mu_{\varepsilon,1}^k$ will necessarily be all contained in $B_{2(R+1)]}(\star_k)$. It follows that each measure $\mu_{\varepsilon,t}^k$ has its support included in $B_{2(R+1)]}(\star_k)$.

From that point on, the very same reasoning as in the proof of Theorem 29.17 can be applied, since, say, the ball $B_{2(R+2)]}(\star_k)$ in $\mathcal{X}_k$ converges in the measured Gromov–Hausdorff topology to the ball $B_{2(R+2)]}(\star)$ in $\mathcal{X}$, etc. □

An application in Riemannian geometry

In this section, by convention I shall say that a metric-measure space (M, d, ν) is a smooth Riemannian manifold if the distance d is the geodesic distance induced by a Riemannian metric g on M, and ν is a reference measure that can be written $e^{-V}\,\mathrm{vol}$, where vol is the volume measure on M and $V \in C^2(M)$. This definition extends in an obvious way to pointed metric-measure spaces. Then Theorem 29.7 guarantees that the synthetic and analytic definitions of $\mathrm{CD}(K,N)$ bounds coincide for Riemannian manifolds.

The next theorem, which is a simple consequence of our previous results, may be seen as one noticeable outcome of the theory of weak $\mathrm{CD}(K,N)$ spaces. Note that it is an external result, in the sense that its statement does not involve the definition of weak $\mathrm{CD}(K,N)$ spaces, nor any reference to optimal transport.

Theorem 29.21 (Smooth limits of $\mathrm{CD}(K,N)$ manifolds are $\mathrm{CD}(K,N)$). *Let $K \in \mathbb{R}$ and $N \in [1,\infty]$. If a sequence of smooth $\mathrm{CD}(K,N)$ manifolds converges to some smooth manifold in the (pointed) measured Gromov–Hausdorff topology, then the limit also satisfies the (classical) $\mathrm{CD}(K,N)$ curvature-dimension bound.*

Proof of Theorem 29.21. The statement follows at once from Theorems 29.7 and 29.18. □

The space of CD(K, N) spaces

Theorem 29.17 can be summarized as follows: *The space of all compact metric-measure spaces satisfying a weak* $\mathrm{CD}(K,N)$ *bound is closed under measured Gromov–Hausdorff convergence.*

In connection with this result, recall Gromov's precompactness theorem (Theorem 27.10): Given $K \in \mathbb{R}$, $N < \infty$ and $D < \infty$, the set $\mathcal{M}(K,N,D)$ of all smooth compact manifolds with dimension bounded above by N, Ricci curvature bounded below by K and diameter bounded above by D is *precompact* in the Gromov–Hausdorff topology. Then it follows from Theorem 29.17 that any element of the closure of $\mathcal{M}(K,N,D)$ is a compact metric-measure length space satisfying $\mathrm{CD}(K,N)$, in the weak sense of Definition 29.6. This conclusion is nontrivial even if the limit space is smooth.

Even if it is smooth, the limit space might have reference measure $\nu = e^{-\Psi}\mathrm{vol}$, for some nonconstant Ψ. Such phenomena do indeed occur in examples where there is a *collapse* in the dimension; that is, when the dimension of the limit manifold is strictly less than the dimension of the manifolds in the converging sequence. The next example shows that basically any reference measure can be obtained as a limite of volume measures of higher-dimensional manifolds; it is a strong motivation to replace the class of Riemannian manifolds by the class of metric-measure spaces.

Example 29.22. Let (M,g) be a compact n-dimensional Riemannian manifold, equipped with its geodesic distance d and its volume vol; let V be any C^2 function on M, and let $\nu(dx) = e^{-V(x)}\, d\mathrm{vol}\,(x)$. Let S^2 stand for the usual 2-dimensional sphere, equipped with its usual metric σ. For $\varepsilon \in (0,1)$, define M_ε to be the warped product of (M,g) by $\varepsilon\, S^2$: This is the $(n+2)$-dimensional manifold $M \times S^2$, equipped with the metric $g_\varepsilon(dx,ds) = g(dx) + \varepsilon^{-2}\, e^{-V(x)}\, \sigma(ds)$. As $\varepsilon \to 0$, M_ε collapses to M; more precisely the manifold $(M_\varepsilon, g_\varepsilon)$, seen as a metric-measure space, converges in measured Gromov–Hausdorff sense to (M,d,ν). Moreover, if $\mathrm{Ric}_{n+2,\nu} \geq K$, then M_ε has Ricci curvature bounded below by K_ε, where $K_\varepsilon \to K$.

We shall see later (Theorem 30.15) that if $(\mathcal{X},d,\nu)$ is a weak $\mathrm{CD}(K,N)$ space, then ν is locally doubling on its support. More precisely, if $\star$ is an arbitrary base point, there is a constant $D = D(K,N,R)$ such that the measure ν is D-doubling on $B_R(\star) \cap \mathrm{Spt}\,\nu$. Combining this with Theorem 27.31, we arrive at the following **compactness theorem**:

Theorem 29.23. *(i) Let $K \in \mathbb{R}$, $N < \infty$ and $D < \infty$, and $0 < m \leq M < \infty$. Let $\mathcal{CDD}(K,N,D,m,M)$ be the space of all compact metric-measure spaces $(\mathcal{X},d,\nu)$ which satisfy the weak curvature-dimension bound $\mathrm{CD}(K,N)$ of Definition 29.6, together with $\mathrm{diam}\,(\mathcal{X},d) \leq D$, $0 < m \leq \nu[\mathcal{X}] \leq M$, and $\mathrm{Spt}\,\nu = \mathcal{X}$. Then $\mathcal{CDD}(K,N,D,m,M)$ is compact in the measured Gromov–Hausdorff topology.*

(ii) Let $K \in \mathbb{R}$, $N < \infty$ and $0 < m \leq M < \infty$. Let $p\mathcal{CDD}(K,N,m,M)$ be the space of all pointed locally compact Polish metric-measure length spaces $(\mathcal{X},d,\nu,\star)$ which satisfy the weak curvature-dimension bound $\mathrm{CD}(K,N)$ of Definition 29.6, $0 < m \leq \nu[B_1(\mathcal{X})] \leq M$, and $\mathrm{Spt}\,\nu = \mathcal{X}$. Then $p\mathcal{CDD}(K,N,m,M)$ is compact in the measured Gromov–Hausdorff topology.

Remark 29.24. It is a natural question whether smooth Riemannian manifolds, equipped with their geodesic distance and their volume measure (multiplied by a positive constant), form a dense set in, say $\mathcal{CDD}(K,N,D,m,M)$. The answer is negative, as will be discussed in the conclusion of these notes.

Bibliographical Notes

Here are some (probably too lengthy) comments about the genesis of Definition 29.6. It comes after a series of particular cases and/or variants studied by Lott and myself [247, 249] on one hand, Sturm [336, 340] on the other hand.

In the case $K = 0$, for compact spaces, Definition 29.6 is exactly the definition that was used in [247], except for the fact that μ_0 and μ_1 were not required to be absolutely continuous with respect to ν, but only to be supported in $\mathrm{Spt}\,\nu$. This difference is irrelevant, because one can always approximate singular measures by absolutely continuous ones, and then pass to the limit in the displacement convexity inequalities; see Theorem 30.8 in the next Chapter. So both definitions are in fact equivalent. My motivation for imposing the absolute continuity in these notes is twofold: (i) in the case of a smooth manifold, absolute continuity of the endpoints measure implies uniqueness of the Wasserstein geodesic; (ii) it simplifies the proof of stability under measured Gromov–Hausdorff convergence.

In the case $N = \infty$, the definition in [247] was about the same as Definition 29.6, but it was based on inequality (29.2) (which is very simple in the case $K = \infty$) instead of (29.3). Sturm [336] also used a similar definition, but preferred to impose the weak displacement convexity inequality only for the Boltzmann H functional, not for the whole class $\mathcal{DC}_\infty$. It is interesting to note that precisely for the H functional and $N = \infty$, inequalities (29.2) and (29.3) are precisely the same, while in general the former is weaker. So the definition which I have adopted here is a priori stronger than both definitions in [247] and [336].

So far only the cases where $N = \infty$ or $K = 0$ had been considered. For the remaining cases, it took some time before a consistent picture emerged. Then Sturm [340] had the brilliant idea to use distorted displacement convexity as the basis for the definition of $\mathrm{CD}(K, N)$ in the finite-dimensional case. His definition is quite close to Definition 29.6, with two differences. First, he does not impose the basic inequality to hold true for all members of the class $\mathcal{DC}_N$, but only for functions of the form $-r^{1-1/N'}$ with $N' \geq N$. Secondly, he does not require the displacement interpolation $(\mu_t)_{0\leq t\leq 1}$ and the coupling π to be related via some dynamical optimal transference plan.

After becoming aware of Sturm's work, Lott and myself [249] modified his definition, requiring the displacement convexity inequality to hold true for all $U \in \mathcal{DC}_N$, imposing a relation between (μ_t) and π, and allowing in addition μ_0 and μ_1 to be singular (provided that their support is included in the support of ν). In the present set of notes, I chose to drop the extension to singular measures, for the same reason alluded to above; then I decided to extend the new definition to the case $N = \infty$.

Sturm [340] proved the stability of his definition under a variant of measured Gromov–Hausdorff convergence, provided that one stays away from the limit Bonnet–Myers diameter. Then Lott and I [249] briefly sketched a proof of stability for our modified definition. Details appear here for the first time.

The treatment of noncompact spaces here is not exactly the same as in [247] or [340]. In the present set of notes I adopted a rather weak point of view in which every "noncompact" statement reduces to the compact case; in particular in Definition 29.6 I only consider compactly supported probability densities. This leads to simpler proofs, but the treatment in [247, Appendix E] is more precise in that it passes to the limit directly in the inequalities for probability measures that are not compactly supported.

Other tentative definitions have been rejected for various reasons. Let me mention three of them:

(i) Imposing the displacement convexity inequality to hold true along *all* displacement interpolations in Definition 29.6, rather than along *some* displacement interpolation. This

concept is not stable under measured Gromov–Hausdorff convergence. (See the last remark in the Conclusion of these notes.)

(ii) Replace the *integrated* displacement convexity inequalities by *pointwise* inequalities, in the style of those appearing in Chapter 14. For instance, with the same notation as in Definition 29.6, one may define

$$\mathcal{J}_t(\gamma_0) := \frac{\rho_0(x)}{\mathbb{E}\left[\rho_t(\gamma_t)|\gamma_0\right]},$$

where γ is a random geodesic with $\text{law}\,(\gamma_t) = \mu_t$, and ρ_t is the absolutely continuous part of μ_t with respect to ν. Then $\mathcal{J}$ is a continuous function of t, and it makes sense to require that inequality (29.1) be satisfied $\nu(dx)$-almost everywhere (as a function of t, in the sense of distributions). This notion of weak $\text{CD}(K,N)$ space makes perfect sense, and is actually stronger than the notion discussed in this chapter. But there is no evidence that it should be stable under measured Gromov–Hausdorff convergence. Integrated convexity inequalities enjoy better stability properties. (One might hope that integrated inequalities lead to pointwise inequalities by a localization argument, as in Chapter 19; but this is not obvious at all, due to the a priori nonuniqueness of displacement interpolation in a nonsmooth context.)

(iii) Choose inequality (29.2) as the basis for the definition, instead of (29.3). In the case $K < 0$, this inequality is stable, due to the convexity of $-r^{1-1/N}$, and the a priori regularity of the speed field provided by Theorem 28.5. (This was actually my original motivation for Theorem 28.5.) In the case $K > 0$ there is no reason to expect that the inequality is stable, but then one can weaken even more the formulation of $\text{CD}(K,N)$ and replace it by

$$U_\nu(\mu_t) \le (1-t)\, U_\nu(\mu_0) + t\, U_\nu(\mu_1) - \frac{K_{N,U}}{2} \max\left(\sup \rho_0, \sup \rho_1\right)^{-1/N} W_2(\mu_0,\mu_1)^2, \quad (29.20)$$

which in turn is stable, and still equivalent to the usual $\text{CD}(K,N)$ when applied to smooth manifolds. For the purpose of the present chapter, this approach would have worked fine; Theorem 29.21 was actually first proved for general K, N by this approach (unpublished work). But basing the definition of the general $\text{CD}(K,N)$ criterion on (29.20) has a strong drawback: it seems very difficult, if not impossible, to derive from it any sharp geometric theorem such as Bishop–Gromov or Bonnet–Myers. We shall see in the next chapter that such sharp inequalities do follow from Definition 29.6.

Now, here are some further comments about the ingredients in the proof of Theorem 29.17.

The extension of U_ν to singular measures (Definition 29.1, Proposition 29.15, Theorem 29.16(i)-(ii)) were worked out in detail in [247]. At least some of these properties belong to folklore, but it is not so easy to find precise references. For the particular case $U(r) = r \log r$, there is a detailed proof of Theorem 29.16(i)-(ii) in [15, Lemmas 9.4.3 to 9.4.5, Corollary 9.4.6] when $\mathcal{X}$ is a separable Hilbert space, possibly infinite-dimensional; the proof of the contraction property in that reference does not rely on the Legendre representation. There is also a proof of the lower semi-continuity and the contraction property, for general functions U, in [240, Chapter 1]; the arguments there do not rely on the Legendre representation either. I personally advocate the use of the Legendre representation, as an efficient and versatile tool.

In [247], we also discussed the extension of these properties to spaces that are not necessarily compact, but only locally compact, and reference measures that are not necessarily

finite, but only locally finite. Integrability conditions at infinity should be imposed on μ, as in Theorem 17.8. The discussion on the Legendre representation in this generalized setting is a bit subtle, for instance it is in general impossible to impose at the same time $\varphi \in C_c(\mathcal{X})$ and $U^*(\varphi) \in C_c(\mathcal{X})$. In these notes, I preferred to limit the use of the Legendre representation to the compact case; but another approximation argument will be used in the next chapter to extend the displacement convexity inequalities to probability measures that are not compactly supported.

The density of $C(\mathcal{X})$ in $L^1(\mathcal{X}, \nu)$, where $\mathcal{X}$ is a compact Hausdorff space and ν is a finite Borel measure, is a classical result that can be found e.g. in [315, Theorem 3.14]. It is also true that nonnegative continuous functions are dense in $L^1_+(\mathcal{X}, \nu)$, or that continuous probability densities are dense in the space of probability densities, equipped with the L^1 norm. All these results can be derived from Lusin's approximation theorem [315, Theorem 2.24]. In [247], Lott and I used a more constructive but more sophisticated approximation procedure, based on regularization kernels. While the scheme of proofs in this chapter is basically the same as in [247], here I avoided the use of regularization kernels by imposing the absolute continuity of μ_0 and μ_1 in Definition 29.6, and using Proposition 29.10 to reduce to Lipschitz continuous nonlinearities: then, according to Theorem 29.16(iii), the associated functionals are continuous on $L^1(\nu)$. Still, regularizing kernels will be useful in the next chapter, when it comes to establish displacement convexity inequalities for possibly singular measures μ_0, μ_1.

Further note that the Lipschitz continuity of U makes it possible to approximate $f(x, y) = \beta(x, y)\, U(\rho(x)/\beta(x, y))$ in $L^1((\mathcal{X}, \nu); C(\mathcal{X}))$ by just $\beta(x, y)\, U(\rho_\varepsilon(x)/\beta(x, y))$, where ρ_ε is a continuous approximation of ρ. For more general nonlinearities (growing "at most polynomially" at infinity), it is still possible to pass to the limit in the displacement convexity inequality, with the help of some vector-valued approximation theorems, such as the density of $C(\mathcal{X} \times \mathcal{X})$ in $L^1((\mathcal{X}, \nu); C(\mathcal{X}))$; but this is quite more painful.[1]

Apart from Theorem 29.21, other "external" consequences of the theory of weak $\mathrm{CD}(K, N)$ spaces are discussed in [247], in the cases $K = 0$ and $N = \infty$.

Lemma 29.5 is taken from a recent work of mine with Figalli [**?**]. I shall use it to prove Theorem 30.30 later in these notes.

Finally, let me give some remarks about the examples considered in this section.

The following generalization of Example 29.11 is proven in [15, Theorems 9.4.10 and 9.4.11]: If ν is a finite measure such that H_ν is displacement convex, then ν takes the form $e^{-V}\mathcal{H}^k$, where V is lower semi-continuous and $\mathcal{H}^k$ is the k-dimensional Hausdorff measure, $k = \dim(\mathrm{Spt}\, \nu)$. The same reference extends to infinite-dimensional separable Hilbert spaces the result according to which H_ν is displacement convex if and only if ν is log-concave.

Example 29.12 was treated by Lott and myself [247]. More precisely, we show that the quotient of a $\mathrm{CD}(K, N)$ Riemannian manifold by a compact Lie group action is still a weak $\mathrm{CD}(K, N)$ space, if $K = 0$ or $N = \infty$. The same theorem is certainly true for all values of K and N, but this was not written down. Elementary background on Lie group actions, and possibly singular spaces obtained by this procedure, can be found in Burago, Burago and Ivanov [81].

Example 29.13 will be considered in more detail in the Conclusion of these notes.

Example 29.22 was explained to me by Lott; it is studied in detail in [**?**].

[1] As a matter of fact, I was working on precisely this problem when my left lung collapsed, earning me a one-week holiday in hospital with unlimited amounts of pain-killers.

30

Weak Ricci curvature bounds II: Geometric and analytic properties

In the previous chapter I introduced the concept of weak curvature-dimension bound, which extends the classical notion of curvature-dimension bound from the world of smooth Riemannian manifolds to the world of metric-measure length spaces; then I proved that such bounds are stable under measured Gromov–Hausdorff convergence.

Still, this notion would be of limited value if it could not be used to derive nontrivial conclusions. But it turns out that weak curvature-dimension bounds can indeed be used to derive interesting geometric and analytic consequences. This might not be a surprise to the reader who has already browsed Part II of these notes, since there many geometric and analytic statements of Riemannian geometry were derived from optimal transport theory.

It will turn out that weak $\mathrm{CD}(K,N)$ spaces satisfy many interesting properties. For instance, in nonbranching weak $\mathrm{CD}(K,N)$ spaces the reference measure satisfies a doubling property and a local Poincaré inequality. So these spaces at least have some kind of regularity; it is a striking manifestation of the "rigidity" associated with lower curvature bounds. (Also the better known Alexandrov spaces with curvature bounded below enjoy remarkable regularity properties.)

So in this last chapter, I shall attempt to present a state of the art about properties of weak $\mathrm{CD}(K,N)$ spaces. This direction of research seems to be growing relatively fast, so the present list might soon become outdated.

Conventions: In all the sequel, a "weak $\mathrm{CD}(K,N)$ space" is a *locally compact, complete separable length space $(\mathcal{X},d)$ equipped with a locally finite measure ν, satisfying a weak $\mathrm{CD}(K,N)$ condition as in Definition 29.6.*

Elementary properties

The next proposition gathers some almost immediate consequences of the definition. I shall say that a subset $\mathcal{X}'$ of a length space $(\mathcal{X},d)$ is *totally convex* if any geodesic whose endpoints belong to $\mathcal{X}'$ is entirely contained in $\mathcal{X}'$.

Proposition 30.1 (Elementary properties of weak $\mathrm{CD}(K,N)$ spaces). *Let $(\mathcal{X},d,\nu)$ be a weak $\mathrm{CD}(K,N)$ space. Then*

(i) If $\mathcal{X}'$ is a totally convex closed subset of $\mathcal{X}$, then $\mathcal{X}'$ inherits from $(\mathcal{X},d,\nu)$ a natural structure of metric-measure length space, and it is also a weak $\mathrm{CD}(K,N)$ space;

(ii) For any $\alpha>0$, the space $(\mathcal{X},d,\alpha\nu)$ is still a weak $\mathrm{CD}(K,N)$ space;

(iii) For any $\lambda>0$, the space $(\mathcal{X},\lambda\, d,\nu)$ is still a weak $\mathrm{CD}(\lambda^{-2}K,N)$ space.

Proof of Proposition 30.1. Property (i) is almost obvious: Equip X' with the restriction of the distance d and the measure ν. Let μ_0, μ_1 be two probability measures in $P_2(\mathcal{X}')$. The notion of optimal coupling is the same whether one considers them as measures on $\mathcal{X}'$ or on $\mathcal{X}$. Also, since $\mathcal{X}'$ is totally convex, a path $[0,1] \to \mathcal{X}$ with endpoints in $\mathcal{X}'$ is a geodesic in $\mathcal{X}'$ if and only if it is a geodesic in $\mathcal{X}$. So $\mathcal{X}'$ is a length space, and the representation theorem for Wasserstein geodesics (Theorem 7.20) ensures that a path $(\mu_t)_{0\le t\le 1}$ valued in $P_2(\mathcal{X}')$ is a geodesic in $P_2(\mathcal{X}')$ if and only if it is a geodesic in $P_2(\mathcal{X})$. Property (i) follows immediately.

To prove (ii), note that the replacement of ν by $\alpha\nu$ induces a multiplication of the density ρ by α^{-1}; so

$$U_{\alpha\nu}(\mu) = (U_\alpha)_\nu(\mu), \qquad U^\beta_{\pi,\alpha\nu}(\mu) = (U_\alpha)^\beta_{\pi,\nu}(\mu),$$

where $U_\alpha(r) = \alpha\, U(\alpha^{-1} r)$. But the transform $U \to U_\alpha$ leaves the class $\mathcal{DC}_N$ invariant. So the inequalities defining the $\mathrm{CD}(K,N)$ condition will hold just the same in $(\mathcal{X}, d, \alpha\nu)$ or in $(\mathcal{X}, d, \nu)$.

As for (iii), recall the definition of $\beta^{(K,N)}$:

$$\beta_t^{(K,N)}(x,y) = \left(\frac{\sin t\alpha(N,K,d(x,y))}{t \sin \alpha(N,K,d(x,y))} \right)^{N-1}, \qquad \alpha(N,K,d) = \sqrt{\frac{K}{N-1}}\, d(x,y).$$

Then $\alpha(N,K,d) = \alpha(N, \lambda^{-2}K, \lambda d)$, from which Property (iii) follows immediately. □

The next theorem shows that the property of being a $\mathrm{CD}(K,N)$ space does not involve the whole space $\mathcal{X}$, but only the support of ν:

Theorem 30.2 (restriction to the support). *A metric-measure space $(\mathcal{X}, d, \nu)$ is a weak $\mathrm{CD}(K,N)$ space if and only if $(\mathrm{Spt}\,\nu, d, \nu)$ is itself a weak $\mathrm{CD}(K,N)$ space.*

Remark 30.3. Theorem 30.2 allows one to systematically reduce to the case when $\mathrm{Spt}\,\nu = \mathcal{X}$ in the study of properties of weak $\mathrm{CD}(K,N)$ spaces. Then why not impose this in the definition of these spaces? The answer is that on some occasions it is useful to allow $\mathcal{X}$ to be larger than $\mathrm{Spt}\,\nu$, in particular for *convergence* issues. Indeed, it may very well happen that a sequence of weak $\mathrm{CD}(K,N)$ spaces $(\mathcal{X}_k, d_k, \nu_k)_{k\in\mathbb{N}}$ with $\mathrm{Spt}\,\nu_k = \mathcal{X}_k$ converges in measured Gromov–Hausdorff sense to a weak $\mathrm{CD}(K,N)$ space $(\mathcal{X}, d, \nu)$ with $\mathrm{Spt}\,\nu \neq \mathcal{X}$. This phenomenon of "reduction of support" is impossible if $N < \infty$, as shown by Theorem 29.23, but can occur in the case $N = \infty$. As a simple example, consider the case when $\mathcal{X}_k = (\mathbb{R}^n, |\cdot|)$ is the Euclidean space $\mathbb{R}^n$ equipped with the sharply peaked Gaussian probability measure $e^{-k|x|^2} dx / Z_k$, where Z_k is a normalizing constant. Then $\mathcal{X}_k$ converges in measured Gromov–Hausdorff sense to $\mathcal{X} = (\mathbb{R}^n, |\cdot|, \delta_0)$. Each of the spaces $\mathcal{X}_k$ is a weak $\mathrm{CD}(0,\infty)$ space and satisfies $\mathrm{Spt}\,\nu_k = \mathcal{X}_k$, however the limit measure is supported in just a point. To summarize things: *For weak $\mathrm{CD}(K,N)$ spaces $(\mathcal{X}, d, \nu)$ with $N < \infty$, one probably does not lose anything by assuming $\mathrm{Spt}\,\nu = \mathcal{X}$; but in the class of weak $\mathrm{CD}(K,\infty)$ spaces, the stability theorem would not be true if one would not allow the support of ν to be strictly smaller than the whole space.*

Proof of Theorem 30.2. First assume that $(\mathrm{Spt}\,\nu, d, \nu)$ is a weak $\mathrm{CD}(K,N)$ space. Replacing $\mathrm{Spt}\,\nu$ by $\mathcal{X}$ does not enlarge the class of absolutely continuous probability measures that can be chosen for μ_0 and μ_1 in Definition 29.6, and does not change the functionals U_ν or $U^{\beta_t^{(K,N)}}_{\pi,\nu}$ either. Because $\mathrm{Spt}\,\nu$ is (by assumption) a length space, geodesics in $\mathrm{Spt}\,\nu$ are also geodesics in $\mathcal{X}$. So geodesics in $P_2(\mathrm{Spt}\,\nu)$ are also geodesics in $P_2(\mathcal{X})$ (it is the converse

that might be false), and then the property of $\mathcal{X}'$ being a weak $\mathrm{CD}(K,N)$ space implies that $\mathcal{X}$ is also a weak $\mathrm{CD}(K,N)$ space.

The converse implication is more subtle. Assume that $(\mathcal{X}, d, \nu)$ is a weak $\mathrm{CD}(K,N)$ space. Let ρ_0, ρ_1 be two compactly supported probability densities on $\mathcal{X}$ (with respect to the reference measure ν). For each $k \in \mathbb{N}$, it is easy to construct *bounded* probability densities $\rho_{k,0}$ and $\rho_{k,1}$ such that $\|\rho_{k,0} - \rho_0\|_{L^1(\nu)} \le k^{-1}$, $\|\rho_{k,1} - \rho_1\|_{L^1(\nu)} \le k^{-1}$. For $t_0 \in \{0,1\}$, let $\mu_{k,t_0} = \rho_{k,t_0}\,\nu$. By the weak $\mathrm{CD}(K,N)$ criterion, for each k there is a Wasserstein geodesic $(\mu_{k,t})_{0\le t\le 1}$ and an associated coupling $\pi_k \in \Pi(\mu,\nu)$ such that for all $t \in [0,1]$ and $U \in \mathcal{DC}_N$,

$$U_\nu(\mu_{k,t}) \le (1-t)\, U^{\beta_{1-t}^{(K,N)}}_{\pi_k,\nu}(\mu_{k,0}) \;+\; t\, U^{\beta_t^{(K,N)}}_{\check{\pi}_k,\nu}(\mu_{k,1}). \tag{30.1}$$

Choosing $H(r) = r\log r$, and using the monotonicity of the $\mathrm{CD}(K,N)$ condition with respect to N, we deduce

$$H_\nu(\mu_{k,t}) \le (1-t)\, H^{\beta_{1-t}^{(K,\infty)}}_{\pi_k,\nu}(\mu_{k,0}) \;+\; t\, H^{\beta_t^{(K,\infty)}}_{\check{\pi}_k,\nu}(\mu_{k,1}). \tag{30.2}$$

By an explicit calculation (as in the proof of (30.9) later in this chapter) the right-hand side is equal to

$$(1-t)\, H_\nu(\mu_{k,0}) \;+\; t\, H_\nu(\mu_{k,1}) - K\,\frac{t(1-t)}{2}\int d(x_0,x_1)^2\,\pi(dx_0\,dx_1),$$

and this quantity is finite since $\mu_{k,0}$, $\mu_{k,1}$ and compactly supported. By (30.2), $H_\nu(\mu_{k,t}) < \infty$ for all $t \in [0,1]$ and for all $k \in \mathbb{N}$. Since $H'(\infty) = \infty$, this implies that $\mu_{k,t}$ is absolutely continuous with respect to ν, and in particular it is supported in $\mathrm{Spt}\,\nu$.

Next, by Ascoli's theorem, there is a subsequence of the family $(\mu_{k,t})$ which converges uniformly in $C([0,1], P_2(\mathcal{X}))$ to some Wasserstein geodesic $(\mu_t)_{0\le t\le 1}$. Since $\mathrm{Spt}\,\nu$ is closed, μ_t is also supported in $\mathrm{Spt}\,\nu$, for each $t \in [0,1]$.

Let (γ_t) be a random geodesic such that $\mu_t = \mathrm{law}\,(\gamma_t)$. The preceding argument shows that $\mathbb{P}\,[\gamma_t \notin \mathrm{Spt}\,\nu] = 0$ for any $t \in [0,1]$. Let $(t_j)_{j\in\mathbb{N}}$ be a dense sequence of times in $[0,1]$, then $\mathbb{P}\,[\exists j;\ \gamma_{t_j} \notin \mathrm{Spt}\,\nu] = 0$. Since γ is continuous and $\mathrm{Spt}\,\nu$ closed, actually $\mathbb{P}\,[\exists t;\ \gamma_t \notin \mathrm{Spt}\,\nu] = 0$. So γ lies entirely in $\mathrm{Spt}\,\nu$, with probability 1. The path $(\mu_t)_{0\le t\le 1}$ is valued in $P_2(\mathrm{Spt}\,\nu)$, and it is a geodesic in the larger space $P_2(\mathcal{X})$; so it is also a geodesic in $P_2(\mathrm{Spt}\,\nu)$.

Then, for any $U \in \mathcal{DC}_N \cap \mathrm{Lip}(\mathbb{R}_+,\mathbb{R}_+)$, we can pass to the limit in (30.1), invoking Theorems 29.16(i) and (iii):

$$\begin{aligned} U_\nu(\mu_t) \le \liminf_{k\to\infty} U_\nu(\mu_{k,t}) \le \limsup_{k\to\infty} \Big[(1-t)\, U^{\beta_{1-t}^{(K,N)}}_{\pi_k,\nu}(\mu_{k,0}) \;+\; t\, U^{\beta_t^{(K,N)}}_{\check{\pi}_k,\nu}(\mu_{k,1})\Big] \\ = \;(1-t)\, U^{\beta_{1-t}^{(K,N)}}_{\pi,\nu}(\mu_0) \;+\; t\, U^{\beta_t^{(K,N)}}_{\check{\pi},\nu}(\mu_1), \end{aligned}$$

where $\mu_0 = \rho_0\,\nu$, $\mu_1 = \rho_1\,\nu$, and π is an optimal coupling between μ_0 and μ_1.

To show that $\mathrm{Spt}\,\nu$ is a weak $\mathrm{CD}(K,N)$ space, it only remains to check that it is a length space; then the conclusion will follow from Proposition 29.10. So let x_0, x_1 be any two points in $\mathrm{Spt}\,\nu$; then for any $r > 0$, we have $\nu[B_r(x_0)] > 0, \nu[B_r(x_1)] > 0$; so it makes sense to define $\mu_0 = 1_{B_r(x_0)}/\nu[B_r(x_0)]$, and $\mu_1 = 1_{B_r(x_1)}/\nu[B_r(x_1)]$. The preceding reasoning shows that there is a random geodesic $\gamma^{(r)}$ which lies entirely in $\mathrm{Spt}\,\nu$, and whose endpoints belong to $B_r(x_0)$ and $B_r(x_1)$. By Ascoli's theorem, there is a subsequence $r_j \to 0$ such that $\gamma^{(r_j)}$ converges uniformly to some random geodesic γ, which necessarily satisfies $\gamma_0 = x_0$, $\gamma_1 = x_1$, and lies entirely in $\mathrm{Spt}\,\nu$. The conclusion is that $\mathrm{Spt}\,\nu$, equipped with the distance d, is indeed a length space. □

Displacement convexity

The definition of weak $\mathrm{CD}(K,N)$ spaces is based upon displacement convexity inequalities, but this inequalities are only required to hold under some restrictions on the initial and final probability measures μ_0 and μ_1, which are required to be absolutely continuous and compactly supported. To exploit the full strength of displacement convexity inequalities, it is important to get rid of these restrictions.

The next theorem shows that the functionals appearing in Definition 29.1 can be extended to measures μ that are not compactly supported, provided that the nonlinearity U belongs to some $\mathcal{DC}_N$ class, and the measure μ admits a moment of order p, where N and p are related through the behavior of ν at infinity.

Theorem 30.4 (Domain of definition of U_ν and $U^\beta_{\pi,\nu}$ for singular measures). *Let $(\mathcal{X}, d)$ be a boundedly compact metric space, equipped with a locally finite Borel measure ν, and let $\star$ be any point in $\mathcal{X}$. Let U be a convex continuous function with $U(0) = 0$, let β be a continuous positive function on $\mathcal{X} \times \mathcal{X}$, and let π be a probability measure on $\mathcal{X} \times \mathcal{X}$. For any Borel measure μ, introduce its Lebesgue decomposition with respect to ν:*

$$\mu = \rho\,\nu + \mu_s.$$

Assume that

$$U \in \mathcal{DC}_N, \qquad \int_{\mathcal{X}} d(\star, x)^p\, \mu(dx) < +\infty,$$

where $N \in [1, \infty]$ and $p \geq 2$ are such that

$$\begin{cases} \displaystyle\int_{\mathcal{X}} \frac{d\nu(x)}{[1 + d(\star, x)]^{p(N-1)}} < +\infty & (N < \infty), \\ \\ \exists c > 0 \qquad \displaystyle\int_M e^{-c\, d(\star,x)^p}\, d\nu(x) < +\infty & (N = \infty). \end{cases} \tag{30.3}$$

If $\mathcal{X}$ is not compact, further assume that β satisfies the following bounds:

$$\begin{cases} \beta \text{ is bounded} & (N < \infty) \\ \\ \displaystyle\int (\log \beta(x,y))_+\, \pi(dx\, dy) < +\infty & (N = \infty). \end{cases} \tag{30.4}$$

Then the following expressions make sense in $\mathbb{R} \cup \{+\infty\}$ and can be taken as generalized definitions of the functionals appearing in Definition 29.1:

$$\begin{cases} U_\nu(\mu) := \displaystyle\int_{\mathcal{X}} U(\rho(x))\, \nu(dx) \;+\; U'(\infty)\, \mu_s[\mathcal{X}], \\ \\ U^\beta_{\pi,\nu}(\mu) := \displaystyle\int_{\mathcal{X}\times\mathcal{X}} U\left(\frac{\rho(x)}{\beta(x,y)}\right) \beta(x,y)\, \pi(dy|x)\, \nu(dx) \;+\; U'(\infty)\, \mu_s[\mathcal{X}]. \end{cases} \tag{30.5}$$

Even if there is no such p, $U^\beta_{\pi,\nu}(\mu)$ still makes sense if $\mu \in P^{\mathrm{ac}}_c(\mathcal{X})$.

Proof of Theorem 30.4. The proof is the same as for Theorem 17.23; there are only two minor differences: (a) ρ is not necessarily a probability density, but still its integral is bounded above by 1; (b) there is an additional term $U'(\infty)\mu_s[\mathcal{X}] \in \mathbb{R} \cup \{+\infty\}$. □

Remark 30.5. Remark 17.24 will also apply here: When one works on weak $\mathrm{CD}(K,N)$ bounds, the restrictions imposed on $\beta = \beta^{(K,N)}$ are automatically satisfied in all cases of interest. (Recall in particular Proposition 29.8.)

The next result is a partial extension of Theorem 29.16. Although it is not the most general result of its kind (see the bibliographical notes for more comments), it will be enough to derive displacement convexity inequalities with a lot of generality. As usual, I shall denote by $M_+(\mathcal{X})$ the set of finite (nonnegative) Borel measures on $\mathcal{X}$, and by $L^1_+(\mathcal{X})$ be the set of nonnegative ν-integrable measurable functions on $\mathcal{X}$.

Theorem 30.6 (lower semi-continuity of U_ν for singular measures). *Let $(\mathcal{X}, d)$ be a boundedly compact metric space, equipped with a locally finite measure ν such that $\mathrm{Spt}\,\nu = \mathcal{X}$. Let $U : \mathbb{R}_+ \to \mathbb{R}_+$ be a Lipschitz continuous convex function, with $U(0) = 0$. Let further $\beta(x,y)$ be a continuous positive function on $\mathcal{X} \times \mathcal{X}$. Then,*

(i) For any $\mu \in M_+(\mathcal{X})$ and any sequence $(\mu_k)_{k\in\mathbb{N}}$ converging weakly to μ in $M_+(\mathcal{X})$,

$$U_\nu(\mu) \le \liminf_{k\to\infty} U_{\nu_k}(\mu_k).$$

(ii) For any probability measure μ, there is a sequence of probability measures $(\mu_k)_{k\in\mathbb{N}}$ converging weakly to μ in $P(\mathcal{X})$, such that (a) each μ_k is absolutely continuous with respect to ν and compactly supported; and (b) for any sequence $(\pi_k)_{k\in\mathbb{N}}$ converging weakly to π in $P(\mathcal{X}\times\mathcal{X})$,

$$\limsup_{k\to\infty} U^\beta_{\pi_k,\nu}(\mu_k) \le U^\beta_{\pi,\nu}(\mu).$$

Remark 30.7. The assumption of Lipschitz continuity on U implies that $U(\rho)$ and $\beta\, U(\rho/\beta)$ are integrable functions, so there is no need to appeal to Theorem 30.4 to make sense of the integrals in (30.5).

Proof of Theorem 30.6.

Let $\star$ be an arbitrary base point, and let $(\chi_R)_{R>0}$ be a $\star$-cutoff as in the Appendix (that is, a family of cutoff continuous functions that are identically equal to 1 on a ball $B_R(\star)$). For any $R > 0$, write

$$U_\nu(\chi_R\,\mu) = \int U(\chi_R\,\rho)\,d\nu + U'(\infty)\int \chi_R\,d\mu_s.$$

Since U is convex nonnegative with $U(0) = 0$, it is nondecreasing; by the monotone convergence theorem,

$$U_\nu(\chi_R\,\mu) \xrightarrow[R\to\infty]{} U_\nu(\mu).$$

In particular,

$$U_\nu(\mu) = \sup_{R>0} U_\nu(\chi_R\,\mu). \tag{30.6}$$

On the other hand, for each fixed R, we have

$$U_\nu(\chi_R\,\mu) = U_{\chi_{R+1}\nu}(\chi_R\,\mu),$$

and then we can apply Proposition 29.15(i) with the compact space $(B_{R+1]}(\star), \nu)$, to get

$$U_\nu(\chi_R\,\mu) = \sup\left\{\int_{\mathcal{X}} \varphi\,\chi_R\,d\mu - \int_{\mathcal{X}} U^*(\varphi)\,\chi_{R+1}\,d\nu;\quad \varphi \in C_b\big(B_{R+1]}(\star)\big),\ 0 < \varphi < U'(\infty)\right\}.$$

The function $\varphi\chi_R$, extended by 0 outside of $B_{R+1]}$, defines a bounded continuous function on the whole of $\mathcal{X}$, so $\mu \longmapsto \int \varphi\,\chi_R\,d\mu$ is continuous with respect to the weak topology of convergence against bounded continuous functions. So $U_\nu(\chi_R\mu)$ is a lower semi-continuous function of μ. This combined with (30.6) shows that U_ν is a lower semi-continuous function of μ, which establishes (i).

Next we turn to property (ii). First note that the proof of Theorem 29.16(iii) extends to the noncompact case; the argument is exactly the same, as soon as one knows that $C_b(\mathcal{X})$ is dense in $L^1(\mathcal{X},\nu)$. (In fact it is true that $C_c(\mathcal{X})$ is dense in $L^1(\mathcal{X},\nu)$; if necessary, consult the Appendix for a refresh about these issues.) So the problem lies in the treatment of possibly singular measures.

Let again $\star$ be an arbitrary base point, and let $(\chi_R)_{R>0}$ be a $\star$-cutoff. For any $R>0$, we can introduce a $(B_{R+1]},\nu)$-regularizing kernel $(K_\varepsilon(x,y))_{0<\varepsilon<1}$ as in the Appendix. Then, for any finite measure μ on $\mathcal{X}$ we define

$$K_\varepsilon(\chi_R\,\mu) = \int_{\mathcal{X}} K_\varepsilon(x,y)\,(\chi_R\,\mu)(dy).$$

If $\mu = \rho\,\nu + \mu_s$ is the Lebesgue decomposition of μ, then

$$K_\varepsilon(\chi_R\,\mu) = \rho_{R,\varepsilon}\nu + K_\varepsilon(\chi_R\,\mu_s), \qquad \rho_{R,\varepsilon}(x) = \int_{\mathcal{X}} K_\varepsilon(x,y)\,\chi_R(y)\,\rho(y)\,\nu(dy).$$

As recalled in the Appendix, $\chi_R\rho$ converges to ρ in $L^1(\nu)$ as $R\to\infty$. Also, for any fixed R, $\rho_{R,\varepsilon}$ converges to $\chi_R\rho$ in $L^1(\nu)$ as $\varepsilon\to 0$. Then we can choose $R_k\to\infty$, and $\varepsilon_k\to 0$ such that

$$\big\|\rho_{R_k,\varepsilon_k} - \rho\big\|_{L^1(\nu)} \xrightarrow[k\to\infty]{} 0.$$

Define

$$\mu_k = \frac{K_{\varepsilon_k}(\chi_{R_k}\mu)}{Z_k},$$

where Z_k is a normalizing constant which converges to 1 as $k\to\infty$. We shall now check that this sequence (μ_k) does the job.

It follows from the definition of the functional $U^\beta_{\pi,\nu}$, the convexity of U and the inequality $U(r)/r \le U'(\infty)$ that (a) $U^\beta_{\pi,\nu}$ is a convex functional of μ; (b) $U^\beta_{\pi,\nu}(\mu) \le U'(\infty)\mu[\mathcal{X}]$. So, for any $\theta\in(0,1)$, we can write

$$\begin{aligned} U^\beta_{\pi_k,\nu}(\mu_k) &\le \theta\,U^\beta_{\pi_k,\nu}\left(\frac{\rho_{R_k,\varepsilon_k}\nu}{Z_k\theta}\right) \;+\; (1-\theta)\,U^\beta_{\pi,\nu}\left(\frac{K_\varepsilon(\chi_{R_k}\,\mu_s)}{Z_k(1-\theta)}\right) \\ &\le \theta\,U^\beta_{\pi_k,\nu}\left(\frac{\rho_{R_k,\varepsilon_k}\nu}{Z_k\theta}\right) \;+\; \frac{U'(\infty)}{Z_k}\,\big(K_\varepsilon(\chi_{R_k}\,\mu_s)\big)[\mathcal{X}]. \end{aligned}$$

In the first term of the right-hand side, we may go to 0 as $\theta\to 1$, using e.g. the continuity of $U^\beta_{\pi_k,\nu}$ on $L^1_+(\nu)$ (or just the monotone convergence theorem, since U is nondecreasing). In the second term, we can use the identity $K_\varepsilon(\chi_R\,\mu_s)[\mathcal{X}] = (\chi_R\,\mu_s)[\mathcal{X}] = \int \chi_R\,d\mu_s$. It follows that

$$U^\beta_{\pi_k,\nu}(\mu_k) \le U^\beta_{\pi_k,\nu}\Big(\frac{\rho_{R_k,\varepsilon_k}}{Z_k}\,\nu\Big) \;+\; \frac{U'(\infty)}{Z_k}\int_{\mathcal{X}} \chi_{R_k}\,d\mu_s. \tag{30.7}$$

Now we can pass to the limit as $k\to\infty$. On one hand, ρ_{R_k,ε_k} converges to ρ in $L^1(\nu)$, and $Z_k\to 1$, so by Theorem 29.16(iii) (extended to the noncompact case),

$$U^\beta_{\pi_k,\nu}(\rho_{R_k,\varepsilon_k}\nu) \xrightarrow[k\to\infty]{} U^\beta_{\pi,\nu}(\rho\,\nu).$$

On the other hand, by dominated convergence,

$$\int_{\mathcal{X}} \chi_{R_k}\, d\mu_s \xrightarrow[k\to\infty]{} \int_{\mathcal{X}} d\mu_s = \mu_s[\mathcal{X}].$$

This combined with (30.7) implies

$$\limsup_{k\to\infty}\, U^\beta_{\pi_k,\nu}(\mu_k) \le U^\beta_{\pi,\nu}(\rho\,\nu) \;+\; U'(\infty)\,\mu_s[\mathcal{X}] \;=\; U^\beta_{\pi,\nu}(\mu).$$

So (ii) is proved. □

The next theorem is the final goal of this section (compare with Definition 29.6):

Theorem 30.8 (generalized displacement convexity inequalities in weak $\mathrm{CD}(K,N)$ spaces). *Let $N \in [1,\infty]$, let $(\mathcal{X}, d, \nu)$ be a weak $\mathrm{CD}(K,N)$ space, and let p satisfy condition (30.3). Let μ_0 and μ_1 be two probability measures in $P_p(\mathcal{X})$, whose supports are included in $\mathrm{Spt}\,\nu$. Then there exists a Wasserstein geodesic $(\mu_t)_{0\le t\le 1}$, and an associated optimal coupling π of (μ_0,μ_1) such that, for all $U \in \mathcal{DC}_N$ and for all $t \in [0,1]$,*

$$U_\nu(\mu_t) \le (1-t)\, U^{\beta^{(K,N)}_{1-t}}_{\pi,\nu}(\mu_0) \;+\; t\, U^{\beta^{(K,N)}_{t}}_{\check\pi,\nu}(\mu_1). \tag{30.8}$$

Furthermore, if $N = \infty$, one also has

$$U_\nu(\mu_t) \le (1-t)\, U_\nu(\mu_0) + t\, U_\nu(\mu_1) - \frac{\lambda(K,U)\, t\,(1-t)}{2}\, W_2(\mu_0,\mu_1)^2, \tag{30.9}$$

where

$$\lambda(K,U) = \inf_{r>0}\, \frac{Kp(r)}{r} \;= \begin{cases} K\,p'(0) & \text{if } K>0, \\ 0 & \text{if } K=0, \\ K\,p'(\infty) & \text{if } K<0. \end{cases} \tag{30.10}$$

These inequalities are the starting point for *all* subsequent inequalities considered in the present chapter.

Remark 30.9. In Definition 29.6, we were careful to impose displacement convexity inequalities along *some* Wasserstein geodesic, because such geodesics might not be unique, due to the lack of smoothness of the space. Now that we are considering possibly singular measures, the nonuniqueness issue is even more acute, and already occurs at the level of smooth manifolds. We shall prove however, as a special case of Theorem 30.26 below, that displacement convexity inequalities on Riemannian manifolds hold true along *all* displacement interpolations, even if the initial and final measures are singular.

Proof of Theorem 30.8. Let μ_0 and μ_1 be as in the statement of the theorem. Use Theorem 30.6(ii) with $\mu = \mu_0$ and then $\mu = \mu_1$, to construct suitable approximating sequences $\mu_{k,0}$ and $\mu_{k,1}$. Since $\mu_{k,0}$ and $\mu_{k,1}$ are compactly supported and absolutely continuous, by definition of the $\mathrm{CD}(K,N)$ criterion there are Wasserstein geodesics $(\mu_{k,t})_{0\le t\le 1}$ and associated couplings π_k of $(\mu_{k,0},\mu_{k,1})$ such that for any $U \in \mathcal{DC}_N$, and for any $t \in [0,1]$,

$$U_\nu(\mu_{k,t}) \le (1-t)\, U^{\beta^{(K,N)}_{1-t}}_{\pi_k,\nu}(\mu_{k,0}) \;+\; t\, U^{\beta^{(K,N)}_{t}}_{\check\pi_k,\nu}(\mu_{k,1}). \tag{30.11}$$

For each k, let Π_k be a dynamical optimal coupling such that $\mu_{k,t} = (e_t)_\# \Pi_k$ and $\pi_k = (e_0,e_1)_\# \Pi_k$. The sets $\{\mu_{k,0};\ k\in\mathbb{N}\}\cup\{\mu_0\}$ and $\{\mu_{k,1};\ k\in\mathbb{N}\}\cup\{\mu_1\}$ are compact for

the weak topology and included in $P_2(\mathcal{X})$; it follows from Corollary 7.20 that the family $(\Pi_k)_{k\in\mathbb{N}}$ converges, up to extraction, to a dynamical optimal transference plan Π with $(e_0)_\#\Pi = \mu_0$ and $(e_1)_\#\Pi = \mu_1$. Then, for each $t \in [0,1]$, $\mu_{k,t}$ converges weakly to μ_t, and also π_k converges weakly to π.

Let us assume that $U \in \mathrm{Lip}(\mathbb{R}_+, \mathbb{R}_+)$. Then by Theorem 30.6(ii), we can pass to the lim sup in the right-hand side of (30.11), and by Theorem 30.6(i), we can pass to the lim inf in the left-hand side of the same equation. So, for any $U \in \mathcal{DC}_N \cap \mathrm{Lip}(\mathbb{R}_+, \mathbb{R}_+)$, and for any $t \in [0,1]$,

$$U_\nu(\mu_t) \le (1-t)\, U^{\beta^{(K,N)}_{1-t}}_{\pi,\nu}(\mu_0) \;+\; t\, U^{\beta^{(K,N)}_{t}}_{\check\pi,\nu}(\mu_1).$$

To establish (30.8), it only remains to remove the assumption $U \in \mathrm{Lip}(\mathbb{R}_+, \mathbb{R}_+)$. This is done exactly as in the proof Proposition 29.10, with only two changes in Step 3 of that proof. First, one should pass to the limit in the singular term, but this is no problem since $U'_\ell(\infty)\,\mu_s[\mathcal{X}] \to U'(\infty)\,\mu_s[\mathcal{X}]$. Secondly, to prove that $U(\rho)$ is integrable, the lower bound $U(\rho(x)) \ge -A\,(\rho(x) + 1_{\rho(x)>0})$ is not enough; instead, one should use the lower bounds for $U(\rho(x))$ and $\beta(x,y)\,U(\rho(x)/\beta(x,y))$ provided by the proof of Theorem 30.4, after noting that the assumptions on the distortion coefficients are always satisfied in view of Remark 17.24.

Now we shall see that (30.9) is a consequence of (30.8). First of all, formula (30.10) follows from the fact that $p(r)/r = r^{1/N}(p(r)/r^{1-1/N})$ is nonincreasing, as a product of nonincreasing nonnegative functions. Obviously, it suffices to show that if π is an optimal coupling of (μ_0,μ_1), then for any $U \in \mathcal{DC}_\infty$,

$$(1-t)\, U^{\beta^{(K,\infty)}_{1-t}}_{\pi,\nu}(\mu_0) \;+\; t\, U^{\beta^{(K,\infty)}_{t}}_{\check\pi,\nu}(\mu_1) \;\le\; (1-t)\, U_\nu(\mu_0) \;+\; t\, U_\nu(\mu_1) \\ -\; \lambda(K,U)\,\frac{t\,(1-t)}{2}\, W_2(\mu_0,\mu_1)^2. \quad (30.12)$$

If $t = 0$, then both sides of (30.12) reduce to $U_\nu(\mu_0)$; similarly, if $t = 1$ both sides reduce to $U_\nu(\mu_1)$. So we may assume that $t \in (0,1)$.

By Proposition 17.7, any $U \in \mathcal{DC}_\infty$ is either linear or satisfies $U'(\infty) > 0$. In the first case, inequality (30.12) is obviously true; in the second case, both sides are $+\infty$ if either μ_0 or μ_1 is singular with respect to ν. So in the sequel we may assume that both μ_0 and μ_1 are absolutely continuous; let ρ_0 and ρ_1 be their respective densities. The case $K = 0$ is trivial, so we may assume $K \neq 0$.

Consider first the case $K > 0$. Since $U \in \mathcal{DC}_\infty$, $\delta \to e^\delta\, U(\rho e^{-\delta})$ is convex. So, for any $\rho \ge 0$ and $\delta > 0$,

$$e^\delta\, U\left(\frac{\rho}{e^\delta}\right) \;\le\; U\left(\frac{\rho}{e^0}\right) e^0 \;-\delta e^\delta p\left(\frac{\rho}{e^\delta}\right) \\ \le U(\rho) \;-\; \lambda(K,U)\,\frac{\delta}{K}\,\rho.$$

Apply this inequality with $\delta = \log \beta^{(K,\infty)}_{1-t}(x_0,x_1) > 0$ and $\rho = \rho_0(x_0)$; then integrate with respect to $\pi(dx_1|x_0)\,\nu(dx_0)$: the result is

$$\begin{aligned} U^{\beta^{(K,\infty)}_{1-t}}_{\pi,\nu}(\mu_0) &\le\; U_\nu(\mu_0) \;-\; \lambda(K,U) \int_{\mathcal{X}\times\mathcal{X}} \left[\frac{\log \beta^{(K,\infty)}_{1-t}(x_0,x_1)}{K}\right] \pi(dx_1|x_0)\,\rho_0(x_0)\,\nu(dx_0) \\ &=\; U_\nu(\mu_0) \;-\; \lambda(K,U) \int_{\mathcal{X}\times\mathcal{X}} \left[\frac{[1-(1-t)^2]\,d(x_0,x_1)^2}{6}\right] \pi(dx_0\,dx_1) \\ &=\; U_\nu(\mu_0) \;-\; \lambda(K,U)\left(\frac{1-(1-t)^2}{6}\right) W_2(\mu_0,\mu_1)^2. \end{aligned} \quad (30.13)$$

Similarly,

$$U_{\pi,\nu}^{\beta_{1-t}^{(K,\infty)}}(\mu_0) \;\le\; U_\nu(\mu_1) \;-\; \lambda(K,U)\left(\frac{1-t^2}{6}\right) W_2(\mu_0,\mu_1)^2. \tag{30.14}$$

Then inequality (30.12) follows from taking the linear combination of (30.13) and (30.14) with respective coefficients $1-t$ and t, and noting that

$$\frac{(1-t)\,[1-(1-t)^2] \;+\; t\,(1-t)^2}{6} \;=\; \frac{t\,(1-t)}{2}.$$

The case $K<0$ is simlar; the change of sign of K is compensated by the change of sign of $\delta = \log \beta^{(K,\infty)}$, so the inequalities are similar. □

Brunn–Minkowski inequality

The next theorem can be taken as the first step in the theory of volume control in weak $\mathrm{CD}(K,N)$ spaces:

Theorem 30.10 (Brunn–Minkowski inequality in weak $\mathrm{CD}(K,N)$ spaces). *Let $K\in\mathbb{R}$ and $N\in[1,\infty]$. Let $(\mathcal{X},d,\nu)$ be a weak $\mathrm{CD}(K,N)$ space, let A_0, A_1 be two compact subsets of $\mathrm{Spt}\,\nu$, and let $t\in(0,1)$. Assume that each of the sets A_0, A_1 either has positive measure or is reduced to a point. Then*

- If $N<\infty$,

$$\nu\big[[A_0,A_1]_t\big]^{\frac{1}{N}} \ge (1-t)\left[\inf_{(x_0,x_1)\in A_0\times A_1} \beta_{1-t}^{(K,N)}(x_0,x_1)^{\frac{1}{N}}\right]\nu[A_0]^{\frac{1}{N}} + \;t\left[\inf_{(x_0,x_1)\in A_0\times A_1} \beta_{t}^{(K,N)}(x_0,x_1)^{\frac{1}{N}}\right]\nu[A_1]^{\frac{1}{N}}. \tag{30.15}$$

- In particular, if $N<\infty$ and $K\ge 0$, then

$$\nu\big[[A_0,A_1]_t\big]^{\frac{1}{N}} \ge (1-t)\,\nu[A_0]^{\frac{1}{N}} \;+\; t\,\nu[A_1]^{\frac{1}{N}}. \tag{30.16}$$

- If $N=\infty$, then

$$\log\frac{1}{\nu\big[[A_0,A_1]_t\big]} \le (1-t)\,\log\frac{1}{\nu[A_0]} \;+\; t\,\log\frac{1}{\nu[A_1]} \;-\; \frac{K\,t(1-t)}{2}\sup_{x_0\in A_0,\,x_1\in A_1} d(x_0,x_1)^2. \tag{30.17}$$

Proof of Theorem 30.10. The proof is the same, *mutatis mutandis*, as the proof of Theorem 18.4: define $\mu_0 = 1_{A_0}/\nu[A_0]$, $\mu_1 = 1_{A_1}/\nu[A_1]$, to be understood as δ_{x_0} (resp. δ_{x_1}) if $A_0=\{x_0\}$ (resp. $A_1=\{x_1\}$); then apply the displacement convexity inequality from Theorem 30.8 with the nonlinearity $U(r) = -r^{1-1/N}$. □

Here below are two interesting corollaries:

Corollary 30.11 (non-atomicity of the support). *Let $K\in\mathbb{R}$ and $N\in[1,\infty]$. If $(\mathcal{X},d,\nu)$ is a weak $\mathrm{CD}(K,N)$ space, then either ν is a Dirac mass, or ν has no atom.*

Corollary 30.12 (Exhaustion by intermediate points). *Let $K \in \mathbb{R}$ and $N \in [1, \infty)$. Let $(\mathcal{X}, d, \nu)$ be a weak $\mathrm{CD}(K, N)$ space, let A be a compact subset of $\mathrm{Spt}\,\nu$ with $\nu[A] > 0$, and let $x \in A$. Then*

$$\nu\big[[x, A]_t\big] \xrightarrow[t \to 1]{} \nu[A].$$

Proof of Corollary 30.11. This corollary will be derived as a consequence of (30.17). By Theorem 30.2, we may assume without loss of generality that $\mathrm{Spt}\,\nu = \mathcal{X}$. Suppose that ν has an atom, i.e. some $x_0 \in \mathcal{X}$ with $\nu[\{x_0\}] > 0$; and yet ν is not the Dirac mass at x_0, so that $\nu[\mathcal{X} \setminus \{x_0\}] > 0$. Define $A_0 = \{x_0\}$ and let A_1 be some compact subset of $\mathcal{X} \setminus \{x_0\}$ such that $\nu[A_1] > 0$. Then for $t > 0$, $[A_0, A_1]_t$ does not contain x_0, but it is included in a ball that shrinks around x_0; it follows that $\nu[[A_0, A_1]_t]$ converges to 0 as $t \to 0$. So $\log(1/\nu[[A_0, A_1]_t]) \to +\infty$ as $t \to 0$; but this contradicts (30.17). □

Proof of Corollary 30.12. The upper bound is easy. Let $R = \max\{d(x, a);\ a \in A\}$. Then $[x, A]_t \subset A_{tR]} = \{y;\ d(y, A) \le tR\}$; so $\nu[[x, A]_t] \le \nu[A_{tR]}]$. By regularity of ν, we can pass to the limit as $t \to 1$ and recover

$$\limsup_{t \to 1} \nu\big[[x, A]_t\big] \le \nu[A]. \tag{30.18}$$

To prove the lower bound, apply (30.15) with $A_0 = \{x\}$, $A_1 = A$. This results in

$$\nu\big[[x, A_1]_t\big]^{\frac{1}{N}} \ge t \inf_{a \in A} \beta_t^{K,N}(x, a)^{\frac{1}{N}}\, \nu[A]^{\frac{1}{N}}.$$

As $t \to 1$, $\inf \beta_t^{(K,N)}(x, a)$ converges to 1, so we may pass to the lim inf and recover

$$\liminf_{t \to 1} \nu\big[[x, A]_t\big] \ge \nu[A].$$

This combines with (30.18) proves the claim. □

Bishop–Gromov inequalities

Once we know that ν has no atom, we can get much more precise information and control on the growth of the volume of balls, and in particular prove sharp Bishop–Gromov inequalities for weak $\mathrm{CD}(K, N)$ spaces with $N < \infty$:

Theorem 30.13 (Bishop–Gromov inequality in metric-measure spaces). *Let $(\mathcal{X}, d, \nu)$ be a weak $\mathrm{CD}(K, N)$ space. Let $x_0 \in \mathcal{X}$ and let $r > 0$ be such that the closed ball $B(x_0, r])$ lies in the support of ν; then $\nu[B(x_0, r])] = \nu[B(x_0, r)]$. Moreover,*

- If $N < \infty$, then for any $x_0 \in \mathrm{Spt}\,\nu$,

$$\frac{\nu[B_r(x_0)]}{\displaystyle\int_0^r s^{(K,N)}(t)\, dt} \qquad \text{is a nonincreasing function of } r, \tag{30.19}$$

where $s^{(K,N)}$ is defined as in Theorem 18.7.

- If $N = \infty$, then for any $x_0 \in \mathrm{Spt}\,\nu$ and any $\delta > 0$, there is a constant $C = C\big(K_-, \delta, \nu[B_\delta(x_0)], \nu[B_{2\delta}(x_0)]$, such that for all $r \ge \delta$,

$$\nu[B_r(x_0)] \le e^{Cr}\, e^{(K_-)\frac{r^2}{2}}; \tag{30.20}$$

$$\nu[B_{r+\delta}(x_0) \setminus B_r(x_0)] \le e^{Cr}\, e^{-K\frac{r^2}{2}} \qquad \text{if } K > 0. \tag{30.21}$$

In particular, if $K > 0$ then

$$\forall K' < K, \qquad \int e^{\frac{K'}{2} d(x_0,x)^2}\, \nu(dx) < +\infty. \tag{30.22}$$

Before providing the proof of this theorem, I shall state two important corollaries. The first one is an exercise for anybody who is familiar with the definitions of the Hausdorff measure and Hausdorff dimension:

Corollary 30.14 (dimension of weak $\mathrm{CD}(K,N)$ spaces). *If $\mathcal{X}$ is a weak $\mathrm{CD}(K,N)$ space with $K \in \mathbb{R}$ and $N \in [1,\infty)$. then the Hausdorff dimension of $\mathrm{Spt}\,\nu$ is at most N.*

The second corollary is the doubling property:

Corollary 30.15 (weak $\mathrm{CD}(K,N)$ spaces are doubling). *If $\mathcal{X}$ is a weak $\mathrm{CD}(K,N)$ space with $K \in \mathbb{R}$, $N < \infty$, $\mathrm{Spt}\,\nu = \mathcal{X}$ and $\mathrm{diam}\,(\mathcal{X}) \le D$, then $(\mathcal{X}, d, \nu)$ is C-doubling with a constant C that depends only on K, N and D.*

Combined with the general theory of Gromov–Hausdorff convergence, as exposed in Chapter 27, this result implies the compactness Theorem 29.23.

Now let us turn to the proof of Theorem 30.13.

Proof of Theorem 30.13. The case where ν is a Dirac mass is trivial and should be treated separately. So by Theorem 30.2, we may assume that ν has no atom.

Next, let $x_0 \in \mathcal{X}$ and $r > 0$. The open ball $B_r(x_0)$ contains $[x, B_{r]}(x_0)]_t$, for all $t \in (0,1)$. By Corollary 30.12,

$$\nu[B_r(x_0)] \ge \lim_{t\to 1} \nu[x, B_{r]}(x_0)]_t = \nu[B_{r]}(x_0)].$$

This shows that $\nu[B_r(x_0)] = \nu[B_{r]}(x_0)]$.

The proof of (30.19) follows the same pattern as the proof of Theorem 18.7. The key ingredient is to apply inequality (30.8) in the case when $\mu_0 = \delta_{x_0}$ and μ_1 is the indicator function of a ball $B_{R]}(x_0)$; and to note that $\nu[\{x_0\}] = 0$. Details are exactly as in Theorem 18.7.

Now let us turn to the case $N = \infty$ and prove (30.20). For brevity I shall write B_r for $B_{r]}(x_0)$. Apply (30.17) with $A_0 = B_\delta$, $A_1 = B_r$, and $t = \delta/(2r) \le 1/2$. For any minimizing geodesic γ going from A_0 to A_1, one has $d(\gamma_0, \gamma_1) \le r + \delta$, so

$$d(x_0, \gamma_t) \le d(x_0, \gamma_0) + d(\gamma_0, \gamma_t) \le \delta + t(r+\delta) \le \delta + 2tr \le 2\delta.$$

So $[A_0, A_1]_t \subset B_{2\delta}$, and it follows from (30.17) that

$$\log \frac{1}{\nu[B_{2\delta}]} \le \Big(1 - \frac{\delta}{2r}\Big)\, \log \frac{1}{\nu[B_\delta]} + \frac{\delta}{2r}\, \log \frac{1}{\nu[B_r]} + \frac{K_-}{2}\, \frac{\delta}{2r}\Big(1 - \frac{\delta}{2r}\Big)\, (r+\varepsilon)^2.$$

This leads to an estimate of the form

$$\nu[B_r] \le \exp\Big(a + br + \frac{c}{r} + \frac{K_-\, r^2}{2}\Big),$$

where a, b, c only depend on δ, $\nu[B_\delta]$ and $\nu[B_{2\delta}]$. The result follows.

The proof of inequality (30.21) is quite the same, with now $A_0 = B_\delta$, $A_1 = B_{r+\delta} \setminus B_r$, $t = \delta/(3r)$.

Finally, to prove (30.22), it suffices to take $\delta = 1$ and write

$$\int e^{\frac{K'}{2}\, d(x_0,x)^2}\, \nu(dx) \le e^{\frac{K'}{2}}\, \nu[B_1] + \sum_{k\ge 1} e^{\frac{K'}{2}\,(k+1)^2} \Big[\int_{B_{k+1}\setminus B_k} \Big]$$
$$\le e^{\frac{K'}{2}} + C e^{C(k+1)} \sum_{k\ge 1} e^{\frac{K'}{2}\,(k+1)^2} e^{-Kk^2} \ < +\infty.$$

□

Uniqueness of geodesics

An important result in Riemannian geometry states that almost all pairs of points in a complete Riemannian manifold are linked by a unique geodesic. This statement *does not* extend to weak $\mathrm{CD}(K,N)$ spaces in general, as will be discussed in the concluding remarks; however, it becomes true if the weak $\mathrm{CD}(K,N)$ criterion is supplemented with a *nonbranching* condition, i.e. any two distinct geodesics cannot coincide on a non-trivial interval:

Theorem 30.16 (Uniqueness of geodesics in nonbranching $\mathrm{CD}(K,N)$ spaces). *Let $(\mathcal{X}, d, \nu)$ be a nonbranching weak $\mathrm{CD}(K,N)$ space with $N < \infty$. Then for $\nu\otimes\nu$-almost all $(x,y) \in \mathcal{X}^2$, there is a unique (constant-speed, minimizing) geodesic joining x to y. More precisely, for any $x \in \mathcal{X}$, the set of points $y \in \mathcal{X}$ which can be joined to x by several geodesics has zero measure.*

Proof. Let $x \in \mathcal{X}$, $r > 0$, $A = B_r(x)$ and $A_t = [x, B_r(x)]_t \subset B_{tr}(x)$. For any $z \in A_t$, there is a geodesic γ joining x to some $y \in Z$, with $\gamma(t) = z$. Assume that there would be another distinct geodesic $\widetilde{\gamma}$ joining x to z; up to a rescaling of time, one may assume that also $\widetilde{\gamma}$ is defined on $[0,t]$, so that $\gamma(0) = x$, $\widetilde{\gamma}(t) = z$. ($\widetilde{\gamma}$ might not be defined after time t.) Then the curve obtained by concatenation of $\widetilde{\gamma}$ on $[0,t]$ and γ on $[t,1]$ is also a geodesic, and it is distinct from $\widetilde{\gamma}$, which is impossible since geodesics are nonbranching. The conclusion is that there is one and only one geodesic joining x to z; it is obtained by reparametrizing the restriction of γ to the interval $[0,t]$.

Let $Z := \cup_{0<t<1} A_t \subset A$. The preceding reasoning shows that if z belongs to Z then there is only one geodesic path joining x to z. The sets A'_t are nonincreasing in t, so

$$\nu\Big[\bigcup_{0<t<1} A_t \Big] = \lim_{t\to 1} \nu\big[[x,A]_t\big] = \nu[A],$$

where the first equality follows from the monotone convergence theorem and the second from Corollary 30.12. This The same can be said of all points z lying in the nondecreasing union of the sets A_t for $0 < t < 1$.

So for any $k \in \mathbb{N}$, the set Z_k of points in $B_k(x)$ which can be joined to x by several geodesics is of zero measure. So the set of points in $\mathcal{X}$ which can be joined to x by several geodesics is contained in the union of all Z_k, and is therefore of zero measure too. □

Regularity of the interpolant

If M is a smooth Riemannian manifold, then the displacement interpolant μ_t between μ_0 and μ_1 is absolutely continuous with respect to the volume measure, as soon as *either* μ_0

or μ_1 is absolutely continuous (Theorem 8.5(ii)). It is not known whether this property is still true in weak $\mathrm{CD}(K,N)$ spaces; but at least some regularity properties are known to be inherited by the displacement interpolant if they are satisfied by *both* μ_0 and μ_1. This is the content of the next theorem.

Theorem 30.17 (absolute continuity of displacement interpolants in weak $\mathrm{CD}(K,N)$ spaces). *Let $(\mathcal{X}, d, \nu)$ be a weak $\mathrm{CD}(K,N)$ space with $K \in \mathbb{R}$ and $N \in [1,\infty)$. Let further μ_0, μ_1 be two probability measures in $P_2(\mathcal{X})$. Then*

(i) Assume that both μ_0 and μ_1 are absolutely continuous with respect to ν; if $K < 0$, further assume that they are compactly supported. Let $(\mu_t)_{0\le t\le 1}$ be a Wasserstein geodesic satisfying the displacement convexity inequalities of Theorem 30.8. Then also μ_t is absolutely continuous, for all $t \in [0,1]$;

(ii) If either μ_0 or μ_1 is absolutely continuous, and $t_0 \in (0,1)$ is given, then one can find a Wasserstein geodesic joining μ_0 to μ_1, such that μ_{t_0} is also absolutely continuous.

(iii) If either μ_0 or μ_1 is not purely singular, then one can find a Wasserstein geodesic joining μ_0 to μ_1, such that for all $t \in [0,1]$, μ_t is not purely singular.

(iv) Assume that $K \ge 0$ and that the respective densities ρ_0 and ρ_1 of μ_0 and μ_1 are bounded. Let $(\mu_t)_{0\le t\le 1}$ be a Wasserstein geodesic satisfying the displacement convexity inequalities of Theorem 30.8. Then the density ρ_t of μ_t is bounded by $\max(\sup\rho_0, \sup\rho_1)$.

Proof of Theorem 30.17. First assume that $K \ge 0$, or, what amounts to the same, $K = 0$. Since μ_0 and μ_1 are absolutely continuous, by the Dunford–Pettis theorem there exists $\Psi : \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$\lim_{r\to\infty} \frac{\Psi(r)}{r} = \infty; \qquad \int \Psi(\rho_0)\, d\nu < +\infty, \qquad \int \Psi(\rho_1)\, d\nu < +\infty.$$

Thanks to Proposition 17.7, one may assume that Ψ belongs to $\mathcal{DC}_N$. Then the convexity inequality

$$\Psi_\nu(\mu_t) \le (1-t)\,\Psi_\nu(\mu_0) + t\,\Psi_\nu(\mu_1) \ < +\infty$$

shows that μ_t is absolutely continuous. This proves (i).

Now consider the case when $K < 0$. It is not hard to prove that in the Dunford–Pettis theorem, one may impose Ψ to have polynomial growth, in the sense that $0 \le U(ar) \le C(a)[U(r)+1]$ for any $a > 1$. Since μ_0 and μ_1 are compactly supported, the distortion coefficients $\beta(x_0,x_1)$ appearing in the right-hand side of the inequalities in Theorem 30.8 are bounded from above and below; then the integrability of $\Psi(\rho)$ also implies the finiteness of $\int \beta(x_0,x_1)\,\Psi(\rho(x_0)/\beta(x_0,x_1))\,\pi(dx_1|x_0)\,\nu(dx_0)$, and the same reasoning as before applies.

To prove (ii), let $U(r) = -Nr^{1-1/N}$; then $U \in \mathcal{DC}_N$ and $U'(\infty) = 0$. Moreover, $U_\nu(\mu) < 0$ as soon as μ is absolutely continuous. Among all dynamical optimal transport plans Π with $(e_0)_\#\Pi = \mu_0$, $(e_1)_\#\Pi = \mu_1$ which satisfy the $\mathrm{CD}(K,N)$ displacement convexity inequality, choose one with minimal $U_\nu(\mu_{t_0})$; there exists one such dynamical optimal transport plan by lower semi-continuity of U_ν and compactness of the set of admissible transference plans.

Assume that μ_{t_0} is not absolutely continuous, i.e. there exists N with $\nu[N] = 0$ and $\mu_{t_0}[N] > 0$. Consider the restricted transport plan Π' obtained by conditioning Π by the event "$\gamma(t_0) \in N$". Then $(e_t)_\#\Pi' = \mu'_t$ satisfies the following properties: (a) $\mu'_t \le \mu_t/\mu_{t_0}[N]$, and in particular μ'_0 is absolutely continuous; (b) μ'_{t_0} is concentrated on N. So $U_\nu(\mu'_{t_0}) = 0$, but $U^{\beta^{(K,N)}_{1-t_0}}_{\pi,\nu}(\mu'_0) < 0$. It follows that the inequality

$$U_\nu(\mu'_{t_0}) \le (1-t_0)\, U^{\beta^{(K,N)}_{1-t_0}}_{\pi,\nu}(\mu'_0) + t_0\, U^{\beta^{(K,N)}_{t_0}}_{\check\pi,\nu}(\mu'_1) \tag{30.23}$$

cannot hold.

On the other hand, there has to be some dynamical optimal transport plan Π'' such that $(e_0)_\# \Pi'' = \mu'_0$, $(e_1)_\# \Pi'' = \mu'_1$ and inequality (30.23) holds true with μ'_{t_0} replaced by $\mu''_{t_0} = (e_{t_0})_\# \Pi''$. In particular, $U_\nu(\mu''_{t_0}) < U_\nu(\mu'_{t_0}) = 0$, which implies that μ''_{t_0} is not purely singular.

Now consider the plan $\widehat{\Pi}$ defined by

$$\widehat{\Pi} = \mathbb{P}\,[\gamma_{t_0} \in N]\, \Pi'' \;+\; 1_{[\gamma_{t_0} \notin N]}\, \Pi. \tag{30.24}$$

This is still a dynamical optimal transport plan, because

$$\begin{aligned}
\int d(\gamma_0,\gamma_1)^2\, \widehat{\Pi}(d\gamma) &= \mathbb{P}\,[\gamma_{t_0} \in N] \int d(\gamma_0,\gamma_1)^2\, \Pi''(d\gamma) + \int 1_{\gamma_{t_0}\notin N} d(\gamma_0,\gamma_1)^2\, \Pi(d\gamma) \\
&= \int d(\gamma_0,\gamma_1)^2\, \Pi'(d\gamma) + \int 1_{\gamma_{t_0}\notin N} d(\gamma_0,\gamma_1)^2\, \Pi(d\gamma) \\
&= \int d(\gamma_0,\gamma_1)^2\, \Pi(d\gamma),
\end{aligned}$$

which is the optimal cost.

It follows from (30.24) and the ν-negligibility of N that

$$\widehat{\rho}_{t_0} = a\, \rho''_{t_0} \;+ 1_{\gamma_{t_0}\notin N} \rho_{t_0} = a\, \rho''_{t_0} + \rho_{t_0},$$

where $\widehat{\rho}_{t_0}$, ρ''_{t_0} and ρ_{t_0} respectively stand for the density of the absolutely continuous parts of $\widehat{\mu}_{t_0}$, μ''_{t_0} and μ_{t_0}, and $a = \mathbb{P}\,[\gamma_{t_0} \in N] > 0$. Then from the minimality property of μ_{t_0},

$$-N \int \rho_{t_0}(x)^{1-1/N}\, d\nu(x) = U_\nu(\mu_{t_0}) \le U_\nu(\widehat{\mu}_{t_0}) = -N \int \big(\rho_{t_0}(x) + a\rho''_{t_0}(x)\big)^{1-1/N}\, d\nu(x).$$

Since a is positive and $r \to r^{1-1/N}$ is strictly increasing, this inequality is possible only if $\rho''_{t_0} = 0$ almost everywhere, but this would be in contradiction with the fact that μ''_{t_0} is not purely singular. The only possibility is that μ_{t_0} is absolutely continuous. This proves (ii).

Statement (iii) is based on the same principle as (ii), but now this is much simpler: Choose $U(r) = -Nr^{1-1/N}$ again, and choose a displacement interpolation (μ_t) satisfying the convexity inequality

$$U_\nu(\mu_t) \le (1-t)\, U^{\beta^{(K,N)}_{1-t}}_{\pi,\nu}(\mu_0) \;+\; t\, U^{\beta^{(K,N)}_{t}}_{\check\pi,\nu}(\mu_1).$$

If, say, μ_0 is not purely singular, then the first term on the right-hand side is negative, while the second one is nonpositive. It follows that $U_\nu(\mu_t) < 0$, and therefore μ_t is not purely singular.

To prove (iv), write ρ_t for the density of μ_t, choose $U(r) = r^p$, which lies in $\mathcal{DC}_N$ for any $p \ge 1$, and deduce that

$$\|\rho_t\|^p_{L^p(\nu)} \le U_\nu(\mu_t) \le (1-t)\, \|\rho_0\|^p_{L^p(\nu)} + t\, \|\rho_1\|^p_{L^p(\nu)} \le \max\big(\|\rho_0\|_{L^p(\nu)}, \|\rho_1\|_{L^p(\nu)}\big)^p. \tag{30.25}$$

Since ρ_0 and ρ_1 belong to $L^1(\nu)$ and $L^\infty(\nu)$, by elementary interpolation they belong to all L^p spaces, so the right-hand side in (30.25) is also finite, and ρ_t belongs to L^p. Then take powers $1/p$ in both sides of (30.25) and pass to the limit as $p \to \infty$, to recover (ii). □

Remark 30.18. The above argument exploited the fact that in the definition of weak $\mathrm{CD}(K,N)$ spaces the displacement convexity inequality is required to hold *for all members of* $\mathcal{DC}_N$ and *along a common Wasserstein geodesic*.

HWI and logarithmic Sobolev inequalities

There is a generalized notion of Fisher information in a metric-measure space $(\mathcal{X}, d, \nu)$:

$$I_\nu(\mu) = \int \frac{|\nabla^- \rho|^2}{\rho}\, d\nu, \qquad \mu = \rho\, \nu,$$

where $|\nabla^- \rho|$ is defined by (20.2) (one may also use $|\nabla \rho|$ in place of $|\nabla^- \rho|$). With this notion, one has the following estimates:

Theorem 30.19 (HWI and logarithmic Sobolev inequalities in weak $\mathrm{CD}(K,\infty)$ spaces). *Let $(\mathcal{X}, d, \nu)$ be a weak $\mathrm{CD}(K,\infty)$ space. Let further μ_0 and μ_1 be two probability measures in $P_2(\mathcal{X})$, such that $\mu_0 = \rho_0\, \nu$ with ρ_0 Lipschitz. Then*

$$H_\nu(\mu_0) \le H_\nu(\mu_1) + W_2(\mu_0, \mu_1) \sqrt{I_\nu(\mu_0)} - \frac{K\, W_2(\mu_0, \mu_1)^2}{2}. \tag{30.26}$$

In particular, if $\nu \in P_2(\mathcal{X})$, then, for any $\mu \in P_2(\mathcal{X})$ with Lipschitz-continuous density,

$$H_\nu(\mu) \le W_2(\mu, \nu) \sqrt{I_\nu(\mu)} - \frac{K\, W_2(\mu, \nu)^2}{2}. \tag{30.27}$$

Consequently, if $K > 0$, then ν satisfies a logarithmic Sobolev inequality with constant K:

$$H_\nu \le \frac{I_\nu}{2K}.$$

Proof of Theorem 30.19. The proof is the same as for Theorems 20.7 and 25.1, *mutatis mutandis*. (The absolute continuity of the displacement interpolant appearing in the proof follows from Theorem 30.17, but also more directly from the finiteness of $H_\nu(\mu_0)$ and $H_\nu(\mu_1)$.) □

Sobolev inequalities

More general versions of the HWI inequality, such as Theorem 20.6, are also available, and then one can derive finite-dimensional Sobolev inequalities. An example is the following statement, whose proof is patterned after the proof of Theorem 21.14:

Theorem 30.20 (Sobolev inequalities in weak $\mathrm{CD}(K,N)$ spaces). *Let $(\mathcal{X}, d, \nu)$ be a weak $\mathrm{CD}(K,N)$ space, where $K < 0$ and $N \in [1, \infty)$. Then, for any $R > 0$ there are constants $A = A(K,N,R)$ and $B = B(K,N,R)$ such that for any Lipschitz function u supported in a ball $B(z,R)$,*

$$\|u\|_{L^{\frac{N}{N-1}}} \le A\, \|\nabla^- u\|_{L^1} + B\, \|u\|_{L^1}. \tag{30.28}$$

On the other hand, it is not known whether weak $\mathrm{CD}(K,N)$ spaces with $K > 0$ and $N < \infty$ satisfy sharp Sobolev inequalities such as (21.9).

Diameter control

Recall from Proposition 29.8 that a weak $\mathrm{CD}(K,N)$ space with $K>0$ and $N<\infty$ satisfies the Bonnet–Myers diameter bound

$$\mathrm{diam}\,(\mathrm{Spt}\,\nu) \le \pi \sqrt{\frac{N-1}{K}}.$$

Slightly weaker conclusions can also be obtained under a priori weaker assumptions: For instance, if $\mathcal{X}$ is at the same time a weak $\mathrm{CD}(0,N)$ space and a weak $\mathrm{CD}(K,\infty)$ space, then there is a universal constant C such that

$$\mathrm{diam}\,(\mathrm{Spt}\,\nu) \le C \sqrt{\frac{N-1}{K}}. \tag{30.29}$$

See the bibliographical notes for more details.

Poincaré inequalities

As was already evoked in Part II of these notes, there are many kinds of Poincaré inequalities, which roughly speaking can be divided into *global* and *local* inequalities. In a nonsmooth context, global Poincaré inequalities can be seen as a replacement for *spectral gap* inequalities for the Laplace operator.

If one does not care about dimension, then there is a general principle (independent of optimal transport) according to which a logarithmic Sobolev inequality with constant K implies a global Poincaré inequality with the same constant; and then from Theorem 30.19 we know that a weak $\mathrm{CD}(K,\infty)$ condition does imply such a logarithmic Sobolev inequality. Now, if one does care about the dimension, then it is possible to adapt the proof of Theorem 21.17 and get from $\mathrm{CD}(K,N)$ the precise Poincaré constant $KN/(N-1)$.

Theorem 30.21 (Global Poincaré inequalities in weak $\mathrm{CD}(K,N)$ spaces). *Let $(\mathcal{X},d,\nu)$ be a weak $\mathrm{CD}(K,N)$ space with $K>0$. Then, for any Lipschitz function $f:\mathcal{X}\to\mathbb{R}$,*

$$\int f\,d\nu = 0 \Longrightarrow \qquad \int f^2\,d\nu \le \left(\frac{N-1}{NK}\right) \int |\nabla^- f|^2\,d\nu,$$

with the convention that $(N-1)/N = 1$ if $N=\infty$.

Local Poincaré inequalities play a key role in the modern geometry of metric spaces, and it is natural to ask whether weak $\mathrm{CD}(K,N)$ spaces satisfy them. For the moment however, the only known result in this direction is for *nonbranching* weak $\mathrm{CD}(K,N)$ spaces (only when N is finite, which is a natural restriction):

Theorem 30.22 (Local Poincaré inequalities in nonbranching weak $\mathrm{CD}(K,N)$ spaces). *Let $K\in\mathbb{R}$, $N\in[1,\infty)$, and let $(\mathcal{X},d,\nu)$ be a nonbranching weak $\mathrm{CD}(K,N)$ space. Let $u:\mathcal{X}\to\mathbb{R}$ be a Lipschitz function, and let $x_0\in\mathcal{X}$. Then for any $R\ge 2r$, there is a constant $P=P(K,N,R)$ such that*

$$\fint_{B_r(x_0)} \Big|u(x) - \langle u\rangle_{B_r(x_0)}\Big|\,d\nu(x) \le\ P(K,N,R)\,r \fint_{B_{2r}(x_0)} |\nabla u|(x)\,d\nu(x), \tag{30.30}$$

where $⨍_B = (\nu[B])^{-1} \int_B$ *is the averaged integral over* B; $\langle u \rangle_B = ⨍_B u \, d\nu$ *is the average of the function* u *on* B; $P(K,N,R) = 2^{2N+1} C(K,N,R) \, D(K,N,R)$, *and the constants* $C(K,N,R)$ *and* $D(K,N,R)$ *are defined by* (19.11) *and* (18.9).

In particular, if $K \geq 0$ *then* $P(K,N,R) = 2^{2N+1}$ *is admissible; so* ν *satisfies a uniform local Poincaré inequality. Moreover,* (30.30) *still holds true if the local "norm of the gradient"* $|\nabla u|$ *is replaced by any upper gradient of* u, *that is a function* g *such that for any Lipschitz path* $\gamma : [0,1] \to \mathcal{X}$,

$$|g(\gamma(0)) - g(\gamma(t))| \leq \int_0^t g(\gamma(t)) \, |\dot\gamma(t)| \, dt.$$

Remark 30.23. It would be desirable to eliminate the nonbranching condition, since it is not always satisfied by weak $\mathrm{CD}(K,N)$ spaces, and rather unnatural in the theory of local Poincaré inequalities.

Proof of Theorem 30.22. Modulo changes of notation, the proof is the same as the proof of Theorem 19.9, once Theorem 30.16 guarantees the almost sure uniqueness of geodesics. □

Talagrand inequalities

With logarithmic Sobolev inequalities come a rich functional apparatus for treating concentration of measure. One may also get concentration from curvature bounds $\mathrm{CD}(K,\infty)$ via Talagrand inequalities. As for the links between logarithmic Sobolev and Talagrand inequalities, they also remain true, at least under more stringent regularity assumptions on $\mathcal{X}$:

Theorem 30.24 (Talagrand and log Sobolev inequalities in measure-metric spaces). *(i) Let* $(\mathcal{X}, d, \nu)$ *be a weak* $\mathrm{CD}(K,\infty)$ *space with* $K > 0$. *Then* ν *lies in* $P_2(\mathcal{X})$ *and satisfies the Talagrand inequality* $T_2(K)$.

(ii) Let $(\mathcal{X}, d, \nu)$ *be a locally compact, Polish length space equipped with a locally finite measure, locally doubling measure* ν, *satisfying a local Poincaré inequality on balls. If* ν *satisfies a logarithmic Sobolev inequality with constant* K, *then also* ν *lies in* $P_2(\mathcal{X})$ *and satisfies the Talagrand inequality* $T_2(K)$.

Remark 30.25. In view of Corollary 30.15 and Theorem 30.22, the regularity assumptions required in (ii) are satisfied if $(\mathcal{X}, d, \nu)$ is a nonbranching weak $\mathrm{CD}(K', N')$ space for some $K' \in \mathbb{R}$, $N' < \infty$; note that the values of K' and N' do not play any role in the conclusion.

Proof of Theorem 30.24. Part (i) is an immediate consequence of (30.22) and (30.26) with $\mu_0 = \nu$.

The proof of part (ii) is similar to the proof of Theorem 22.15, once one has an analogue of Proposition 22.13. It turns out that properties (i)-(vi) of Proposition 22.13 are still satisfied when the Riemannian manifold M is replaced by any metric space $\mathcal{X}$, but property (vii) might fail in general. Still it is true that this property holds true *for* ν*-almost all* x, under the assumption that ν is locally doubling and satisfies a local Poincaré inequality on balls. (See the bibliographical notes for references.) This is enough for the proof of Theorem 22.15 to go through. □

There are also dimensional versions of Talagrand inequalities available, for instance the analogue of Theorem 22.34 holds true in weak $\mathrm{CD}(K,N)$ spaces with $K > 0$ and $N < \infty$.

Equivalence of definitions in nonbranching spaces

In the definition of weak $\mathrm{CD}(K,N)$ spaces we chose to impose the displacement convexity inequality for all $U \in \mathcal{DC}_N$, but only along some displacement interpolation. We could have chosen otherwise, for instance impose the inequality for just some particular functions U, or along all displacement interpolations. In the end our choice was dictated partly by the will to get a stable definition, partly by convenience. It turns out that in *nonbranching* metric-measure spaces, the choice really does not matter: It is equivalent

- to require the displacement convexity inequality to hold true for any $U \in \mathcal{DC}_N$; or just for $U = U_N$, where $U_N(r) = -N r^{1-1-/N}$ if $1 < N < \infty$, and $U_\infty(r) = r \log r$;

- to require the inequality to hold true for compactly supported, absolutely continuous probability measures μ_0, μ_1; or for any two probability measures with suitable moment conditions;

- to require the inequality to hold true along some displacement interpolation, or along any displacement interpolation.

The next statement makes this claim precise. Note that I leave apart the case $N = 1$ which is special (for instance U_1 is not defined). I shall write $(U_N)_\nu = H_{N,\nu}$, and $(U_N)^\beta_{\pi,\nu} = H^\beta_{N,\pi,\nu}$.

Theorem 30.26 (Equivalence of definitions in nonbranching spaces).
Let $(\mathcal{X}, d, \nu)$ be a nonbranching locally compact Polish length space equipped with a locally finite measure ν. Let $K \in \mathbb{R}$, $N \in (1, \infty]$, and let $p \in [2, +\infty) \cup \{c\}$ satisfy the assumptions of Theorem 30.4. Then the following three properties are equivalent:

(i) $(\mathcal{X}, d, \nu)$ is a weak $\mathrm{CD}(K,N)$ space, in the sense of Definition 29.6;

(ii) For any two compactly supported probability densities ρ_0 and ρ_1, there is a displacement interpolation $(\mu_t)_{0 \le t \le 1}$ joining $\mu_0 = \rho_0 \, \nu$ to $\mu_1 = \rho_1 \, \nu$, and an associated optimal plan π, such that for all $t \in [0,1]$,

$$H_{N,\nu}(\mu_t) \le (1-t) \, H^{\beta^{(K,N)}_{1-t}}_{N,\pi,\nu}(\mu_0) \; + \; t \, H^{\beta^{(K,N)}_{t}}_{N,\pi,\nu}(\mu_1). \tag{30.31}$$

(iii) For any displacement interpolation $(\mu_t)_{0 \le t \le 1}$ with $\mu_0, \mu_1 \in P_p(\mathcal{X})$, for any associated transport plan π, for any $U \in \mathcal{DC}_N$ and for any $t \in [0,1]$,

$$U_\nu(\mu_t) \le (1-t) \, U^{\beta^{(K,N)}_{1-t}}_{\pi,\nu}(\mu_0) \; + \; t \, U^{\beta^{(K,N)}_{t}}_{\check{\pi},\nu}(\mu_1). \tag{30.32}$$

Theorem 30.26 is interesting even for smooth Riemannian manifolds, since it covers singular measures, for which there is a priori no uniqueness of displacement interpolant. Its proof is based on the idea, already used in Theorem 19.2, that we may condition the optimal transport to lie in a very small ball at time t, and, by passing to the limit, retrieve a pointwise control of the density ρ_t. This will work because the nonbranching property implies the uniqueness of the displacement interpolation between intermediate times, and forbids the crossing of geodesics used in the optimal transport, as in Theorem 7.27. Apart from this simple idea, the proof is quite technical and the reader might skip it at first reading.

Proof of Theorem 30.26. Let us consider for instance the case $N < \infty$; the case $N = \infty$ can be treated in a similar way.

Clearly, (iii) $\Rightarrow$ (i) $\Rightarrow$ (ii). So it is sufficient to show that (ii) $\Rightarrow$ (iii). In the sequel, I shall assume that Property (ii) is satisfied. By the same arguments as in the proof of

Theorem 30.8, it is sufficient to establish (30.32) when U is nonnegative and Lipschitz continuous, and $u(r) := U(r)/r$ is a continuous function of r (with $u(0) = U'(0)$). I shall fix $t \in (0,1)$ and establish Property (iii) for that t. For simplicity I shall abbreviate $\beta_t^{(K,N)}$ into just β_t.

First of all, let us establish that Property (ii) also applies if μ_0 and μ_1 are not absolutely continuous. The scheme of reasoning is the same as we already used several times. Let μ_0 and μ_1 be any two compactly supported measures. Construct (for instance by means of regularizing kernels as in the Appendix) absolutely continuous probability measures $\mu_{k,0}$ and $\mu_{k,1}$ such that $\mu_{k,0}$ converges to μ_0 in $P(\mathcal{X})$ and $\mu_{k,1}$ converges to μ_1 in $P(\mathcal{X})$, all these measures being supported in a uniform compact set. For each $k \in \mathbb{N}$, there exists a dynamical optimal transference plan Π_k such that

$$H_{N,\nu}(\mu_{k,t}) \le (1-t)\, H^{\beta_{1-t}}_{N,\pi_k,\nu}(\mu_{k,0}) \;+\; t\, H^{\beta_t}_{N,\check{\pi}_k,\nu}(\mu_{k,1}), \tag{30.33}$$

where $\mu_{k,t} := (e_t)_\# \Pi_k$ and $\pi_k := (e_0, e_1)_\# \Pi_k$. Since all the measures $\mu_{k,0}$ and $\mu_{k,1}$ are supported in a uniform compact set, Corollary 7.20 guarantees that the sequence $(\Pi_k)_{k\in\mathbb{N}}$ converges, up to extraction, to some dynamical optimal transference plan Π with $(e_0)_\# \Pi = \mu_0$ and $(e_1)_\# \Pi = \mu_1$. Then for all $t \in [0,1]$, $\mu_{k,t}$ converges weakly to $\mu_t = (e_t)_\# \Pi$, and $\pi_k := (e_0, e_1)_\# \Pi_k$ converges weakly to $\pi = (e_0, e_1)_\# \Pi$.

Since the $\mu_{k,t}$ lie in a uniform compact, we can use the lower semi-continuity of U_ν for $U = U_N$ (Theorem 29.16(i)) to get

$$H_{N,\nu}(\mu_t) \le \liminf_{k\to\infty} H_{N,\nu}(\mu_{k,t}). \tag{30.34}$$

It remains to pass to the lim sup in the right-hand side of (30.33). We cannot any longer use Theorem 29.16(iii), since $U = U_N$ is not Lipschitz continuous. On the other hand (by Proposition 17.7 or by more elementary means) we can find a nonincreasing sequence of Lipschitz functions $U^{(m)}$ ($m \in \mathbb{N}$) such that $U^{(m)}$ converges pointwise to $U = U_N$ as $m \to \infty$. Then, for each given m, by Theorem 29.16(iii),

$$U^{\beta_{1-t}}_{\pi_k,\nu}(\mu_{k,0}) \le (U^{(m)})^{\beta_{1-t}}_{\pi_k,\nu}(\mu_{k,0}) \xrightarrow[k\to\infty]{} (U^{(m)})^{\beta_{1-t}}_{\pi,\nu}(\mu_0). \tag{30.35}$$

On the other hand, by dominated convergence,

$$(U^{(m)})^{\beta_{1-t}}_{\pi,\nu}(\mu_0) \xrightarrow[m\to\infty]{} U^{\beta_{1-t}}_{\pi,\nu}(\mu_0). \tag{30.36}$$

It follows from (30.35) and (30.36) that

$$\limsup_{k\to\infty}\; U^{\beta_{1-t}}_{\pi_k,\nu_k}(\mu_{k,0}) \le U^{\beta_{1-t}}_{\pi,\nu}(\mu_0).$$

Of course a similar inequality holds with time $t = 0$ replaced by $t = 1$, and combining this with (30.34) we recover

$$H_{N,\nu}(\mu_t) \le (1-t)\, H^{\beta_{1-t}}_{N,\pi,\nu}(\mu_0) \;+\; t\, H^{\beta_t}_{N,\check{\pi},\nu}(\mu_1).$$

So the statement in Property (ii) remains true even if μ_0 and μ_1 are not absolutely continuous.

Next, the proofs of Theorem 30.13 and Corollary 30.15 goes through, since it only uses the convex function $U = U_N$; in particular the measure ν is locally doubling on its support. This information will be crucial in the sequel.

Also the proof of Theorem 30.17(ii) and (iii) can be easily adapted in the present setting, as soon as μ_0 and μ_1 are compactly supported.

Now we can start the core of the argument. It will be decomposed in four steps.

Step 1: Assume that μ_0 *and* μ_1 *are compactly supported,* μ_t *is absolutely continuous and there exists a dynamical optimal transference plan* Π *joining* μ_0 *to* μ_1*, such that for any subplan* $\Pi' = \widetilde{\Pi}/\widetilde{\Pi}[\Gamma]$, $0 \leq \widetilde{\Pi} \leq \Pi$*, it happens that* Π' *is the unique dynamical optimal transference plan between* $\mu_0' = (e_0)_\# \Pi'$ *and* $\mu_1' = (e_1)_\# \Pi'$*.*

In particular, Π is the unique dynamical optimal transference plan between μ_0 and μ_1, and, by Corollary 7.20, $\mu_t = (e_t)_\# \Pi$ defines the unique displacement interpolation between μ_0 and μ_1. In the sequel, I shall denote by ρ_t the density of μ_t, and by ρ_0, ρ_1 the densities of the absolutely continuous parts of μ_0, μ_1 respectively. I shall also fix Borel sets S_0, S_1 such that $\nu[S_0] = \nu[S_1] = 0$, $\mu_{0,s}$ is concentrated on S_0 and $\mu_{1,s}$ is concentrated on S_1. By convention ρ_0 is defined to be $+\infty$ on S_0, and similarly ρ_1 is defined to be $+\infty$ on S_1.

Let then $y \in \mathrm{Spt}\, \mu_t$, and let $\delta > 0$. Define

$$\mathcal{Z} = \Big\{ \gamma \in \Gamma;\ \gamma_t \in B_\delta(y) \Big\},$$

and let $\Pi' = (1_{\mathcal{Z}} \Pi)/\Pi[\mathcal{Z}]$. (If γ is a random variable distributed according to Π, then Π' is the law of γ conditioned by the event "$\gamma_t \in B_\delta(y)$".) Let $\mu_t' = (e_t)_\# \Pi'$, let ρ_t' be the density of the absolutely continuous part of μ_t', and let $\pi' := (e_0, e_1)_\# \Pi'$. Since Π' is the unique dynamical optimal transference plan between μ_0' and μ_1', we can write the displacement convexity inequality

$$H_{N,\nu}(\mu_t') \leq (1-t)\, H^{\beta_{1-t}}_{N,\pi,\nu}(\mu_0') \ + \ t\, H^{\beta_t}_{N,\check{\pi},\nu}(\mu_1').$$

In other words,

$$\int_{\mathcal{X}} (\rho_t')^{1-\frac{1}{N}}\, d\nu \geq (1-t) \int_{\mathcal{X}\times\mathcal{X}} (\rho_0'(x_0))^{-\frac{1}{N}} \beta_{1-t}(x_0,x_1)^{\frac{1}{N}}\, \pi'(dx_0\, dx_1) + \ t \int_{\mathcal{X}\times\mathcal{X}} (\rho_1'(x_1))^{-\frac{1}{N}} \beta_t(x_0,x_1)^{\frac{1}{N}}\, \pi'(dx_0\, dx_1), \quad (30.37)$$

with the convention that $\rho_0'(x_0) = +\infty$ when $x_0 \in S_0$, and $\rho_1'(x_1) = +\infty$ when $x_1 \in S_1$.

By reasoning as in the proof of Theorem 19.2, we obtain

$$\frac{\nu[B_\delta(y)]^{\frac{1}{N}}}{\mu_t[B_\delta(y)]^{\frac{1}{N}}} \geq \mathbb{E}_{\,\Pi}\Big[(1-t) \left(\frac{\beta_{1-t}(\gamma_0,\gamma_1)}{\rho_0(\gamma_0)} \right)^{\frac{1}{N}} \ + \ t \left(\frac{\beta_t(\gamma_0,\gamma_1)}{\rho_1(\gamma_1)} \right)^{\frac{1}{N}} \Big|\ \gamma_t \in B_\delta(y) \Big].$$

If we define

$$f(\gamma) := (1-t) \left(\frac{\beta_{1-t}(\gamma_0,\gamma_1)}{\rho_0(\gamma_0)} \right)^{\frac{1}{N}} \ + \ t \left(\frac{\beta_t(\gamma_0,\gamma_1)}{\rho_1(\gamma_1)} \right)^{\frac{1}{N}},$$

then the conclusion can be rewritten

$$\frac{\nu[B_\delta(y)]^{\frac{1}{N}}}{\mu_t[B_\delta(y)]^{\frac{1}{N}}} \geq \mathbb{E}_{\,\Pi}[f(\gamma) | \gamma_t \in B_\delta(y)] = \frac{\mathbb{E}\, f(\gamma) 1_{[\gamma_t \in B_\delta(y)]}}{\mu_t[B_\delta(y)]}. \quad (30.38)$$

In view of the nonbranching property, Π only sees geodesics which do not cross each other; recall Theorem 7.27(iv). Let F_t be the map appearing in that theorem, defined by $F_t(\gamma_t) = \gamma$. Then (30.38) becomes

$$\frac{\nu[B_\delta(y)]^{\frac{1}{N}}}{\mu_t[B_\delta(y)]^{\frac{1}{N}}} \geq \frac{\mathbb{E}\,\left[f(F_t(\gamma_t))\,1_{[\gamma_t \in B_\delta(y)]}\right]}{\mu_t[B_\delta(y)]}$$
$$= \frac{\displaystyle\int_{B_\delta(y)} f(F_t(x))\,d\mu_t(x)}{\mu_t[B_\delta(y)]}. \tag{30.39}$$

Since the measure ν is locally doubling, we can apply Lebesgue's density theorem: There is a set N of zero ν-measure such that if $y \notin N$, then

$$\frac{\nu[B_\delta(y)]^{\frac{1}{N}}}{\mu_t[B_\delta(y)]^{\frac{1}{N}}} \quad \xrightarrow[\delta \to 0]{} \quad \frac{1}{\rho_t(y)^{\frac{1}{N}}}.$$

Similarly, outside of a set of zero measure,

$$\frac{\displaystyle\int_{B_\delta(y)} f(F_t(x))\,d\mu_t(x)}{\mu_t[B_\delta(y)]} = \frac{\displaystyle\int_{B_\delta(y)} f(F_t(x))\,\rho_t(x)\,d\nu(x)}{\nu[B_\delta(y)]}\,\frac{\nu[B_\delta(y)]}{\mu_t[B_\delta(y)]} \quad \xrightarrow[\delta \to 0]{} \quad \frac{f(F_t(y))\,\rho_t(y)}{\rho_t(y)},$$

and this coincides with $f(F_t(y))$ if $\rho_t(y) \neq 0$. All in all, $\mu_t(dy)$-almost surely,

$$\frac{1}{\rho_t(y)^{\frac{1}{N}}} \geq f(F_t(y)).$$

Equivalently, $\Pi(d\gamma)$-almost surely,

$$\frac{1}{\rho_t(\gamma_t)^{\frac{1}{N}}} \geq f(F_t(\gamma_t)) = f(\gamma).$$

Let us recapitulate: We have shown that $\Pi(d\gamma)$-almost surely,

$$\frac{1}{\rho_t(\gamma_t)^{\frac{1}{N}}} \geq (1-t)\,\left(\frac{\beta_{1-t}(\gamma_0,\gamma_1)}{\rho_0(\gamma_0)}\right)^{\frac{1}{N}} \;+\; t\,\left(\frac{\beta_t(\gamma_0,\gamma_1)}{\rho_1(\gamma_1)}\right)^{\frac{1}{N}}. \tag{30.40}$$

In the case $N = \infty$, this inequality should be modified into

$$\log \frac{1}{\rho_t(\gamma_t)} \geq (1-t)\,\log \frac{1}{\rho_0(\gamma_0)} + t\,\log \frac{1}{\rho_1(\gamma_1)} + \frac{K\,t(1-t)}{2}\,d(\gamma_0,\gamma_1)^2. \tag{30.41}$$

Step 2: Now we shall prove inequality (30.32) in the case when μ_0 *and* μ_1 *are compactly supported, and* μ_t *is absolutely continuous.* So let $(\mu_s)_{0 \leq s \leq 1}$ be a displacement interpolation joining μ_0 to μ_1, and let Π be a dynamical optimal transport plan with $\mu_s = (e_s)_\# \Pi$. Let $\varepsilon \in (0,1)$ be given. By the nonbranching property and Theorem 7.27, the restricted plan $\Pi^{0,1-\varepsilon}$ obtained by taking the push-forward of Π under the restriction map from $C([0,1];\mathcal{X})$ to $C([0,1-\varepsilon];\mathcal{X})$ is the only dynamical optimal transport plan between μ_0 and $\mu_{1-\varepsilon}$; and more generally, if $0 \leq \widetilde{\Pi} \leq \Pi^{0,1-\varepsilon}$ with $\widetilde{\Pi}[\Gamma] > 0$, then $\Pi' := \widetilde{\Pi}/\widetilde{\Pi}[\Gamma]$ is the only dynamical optimal transport plan between its endpoints measures. In other words, $\widetilde{\mu}_0 = \mu_0$ and $\widetilde{\mu}_1 = \mu_{1-\varepsilon}$ satisfy the assumptions used in Step 1. The only displacement interpolation between $\widetilde{\mu}_0$ and $\widetilde{\mu}_1$ is $\widetilde{\mu}_t = \mu_{(1-\varepsilon)t}$, so we can apply formula (30.40) to that path, after time-reparametrization. Writing

$$t = \left(\frac{1-t-\varepsilon}{1-\varepsilon}\right) \times 0 \;+\; \left(\frac{t}{1-\varepsilon}\right) \times (1-\varepsilon),$$

we see that, $\Pi(d\gamma)$-almost surely,

$$\frac{1}{\rho_t(\gamma_t)^{\frac{1}{N}}} \geq \left(\frac{1-t-\varepsilon}{1-\varepsilon}\right) \left(\frac{\beta_{\frac{1-t-\varepsilon}{1-\varepsilon}}(\gamma_0,\gamma_{1-\varepsilon})}{\rho_0(\gamma_0)}\right)^{\frac{1}{N}} + \left(\frac{t}{1-\varepsilon}\right) \left(\frac{\beta_{\frac{t}{1-\varepsilon}}(\gamma_0,\gamma_{1-\varepsilon})}{\rho_{1-\varepsilon}(\gamma_{1-\varepsilon})}\right)^{\frac{1}{N}}. \quad (30.42)$$

Next, we apply the same reasoning on the time-interval $[t,1]$ rather than $[0,1-\varepsilon]$. We write $1-\varepsilon$ as an intermediate point between t and 1:

$$1-\varepsilon = \left(\frac{\varepsilon}{1-t}\right) \times t \; + \; \left(\frac{1-t-\varepsilon}{1-t}\right) \times 1.$$

Since μ_t is absolutely continuous and $\mu_{1-\varepsilon}$ belongs to the unique displacement interpolation between μ_t and μ_1, it follows from Theorem 30.17(ii) that $\mu_{1-\varepsilon}$ is absolutely continuous too. Then formula (30.40) becomes, after time-reparametrization,

$$\frac{1}{\rho_{1-\varepsilon}(\gamma_{1-\varepsilon})^{\frac{1}{N}}} \geq \left(\frac{\varepsilon}{1-t}\right) \left(\frac{\beta_{\frac{\varepsilon}{1-t}}(\gamma_t,\gamma_1)}{\rho_t(\gamma_t)}\right)^{\frac{1}{N}} + \left(\frac{1-t-\varepsilon}{1-t}\right) \left(\frac{\beta_{\frac{1-t-\varepsilon}{1-t}}(\gamma_t,\gamma_1)}{\rho_1(\gamma_1)}\right)^{\frac{1}{N}}. \quad (30.43)$$

The combination of (30.42) and (30.43) yields

$$\left(1 - \left(\frac{t}{1-\varepsilon}\right)\left(\frac{\varepsilon}{1-t}\right) \beta_{\frac{t}{1-\varepsilon}}(\gamma_0,\gamma_{1-\varepsilon})^{\frac{1}{N}}\, \beta_{\frac{\varepsilon}{1-t}}(\gamma_t,\gamma_1)^{\frac{1}{N}}\right) \frac{1}{\rho_t(\gamma_t)^{\frac{1}{N}}} \geq$$

$$\left(\frac{1-t-\varepsilon}{1-\varepsilon}\right) \left(\frac{\beta_{\frac{1-t-\varepsilon}{1-\varepsilon}}(\gamma_0,\gamma_{1-\varepsilon})}{\rho_0(\gamma_0)}\right)^{\frac{1}{N}} + \left(\frac{1-t-\varepsilon}{1-t}\right)\left(\frac{t}{1-\varepsilon}\right) \left(\frac{\beta_{\frac{1-t-\varepsilon}{1-t}}(\gamma_t,\gamma_{1-\varepsilon})\, \beta_{\frac{t}{1-\varepsilon}}(\gamma_0,\gamma_{1-\varepsilon})}{\rho_1(\gamma_1)}\right)^{\frac{1}{N}}.$$

Then we can pass to the limit as $\varepsilon \to 0$ thanks to the continuity of γ and β; since $\beta_1(x,y) = 1$ for all x, y, we obtain that inequality (30.40) holds true almost surely.

Now let $w(\delta) = u(\delta^{-N}) = \delta^N U(\delta^{-N})$, with the convention $w(0) = U'(\infty)$. By assumption w is a convex nonincreasing function of δ. So

$$\mathbb{E}\, u(\rho_t(\gamma_t)) = \mathbb{E}\, w\left(\frac{1}{\rho_t(\gamma_t)^{\frac{1}{N}}}\right) \leq (1-t)\, \mathbb{E}\, w\left(\left(\frac{\beta_{1-t}(\gamma_0,\gamma_1)}{\rho_0(\gamma_0)}\right)^{\frac{1}{N}}\right) + \; t\, \mathbb{E}\, w\left(\left(\frac{\beta_t(\gamma_0,\gamma_1)}{\rho_1(\gamma_1)}\right)^{\frac{1}{N}}\right). \quad (30.44)$$

The left-hand side of (30.44) is just $\int U(\rho_t(x))/\rho_t(x)\, d\mu_t(x) = \int U(\rho_t(x))\, d\nu(x) = U_\nu(\mu_t)$. The first term in the right-hand side of (30.44) is $(1-t)\, U^{\beta_{1-t}}_{\pi,\nu}(\mu_0)$, since we chose to define $\rho_0(x_0) = +\infty$ when x_0 belongs to the singular set S_0. Similarly, the second term is $t\, U^{\beta_t}_{\check{\pi},\nu}(\mu_1)$. So (30.44) reads

$$U_\nu(\mu_t) \leq (1-t)\, U^{\beta_{1-t}}_{\pi,\nu}(\mu_0) \; + \; t\, U^{\beta_t}_{\check{\pi},\nu}(\mu_1),$$

as desired.

Step 3: Now we wish to establish inequality (30.32) in the case when μ_t *is absolutely continuous*, that is, we just want to drop the assumption of compact support.

It follows from Step 2 that $(\mathcal{X}, d, \nu)$ is a weak $\mathrm{CD}(K,N)$ space, so we now have access to Theorem 30.17 even if μ_0 and μ_1 are not compactly supported; and also we can appeal to Theorem 30.8 to guarantee that Property (ii) is verified for probability measures that

are not necessarily compactly supported. Then we can repeat Steps 1 and 2 without the assumption of compact support, and in the end establish inequality (30.32) for measures that are not compactly supported.

Step 4: Now we shall consider the case when μ_t is not absolutely continuous. (This is the part of the proof which has interest even in a smooth setting.) Let $(\mu_t)_s$ stand for the singular part of μ_t, and $m := (\mu_t)_s[\mathcal{X}] > 0$.

Let $E^{(a)}$ and $E^{(s)}$ be two disjoint Borel sets in $\mathcal{X}$ such that the absolutely continuous part of μ_t is concentrated on $E^{(a)}$, and the singular part of μ_t is concentrated on $E^{(s)}$. Obviously, $\Pi[\gamma_t \in E^{(s)}] = (\mu_t)_s[\mathcal{X}] = m$, and $\Pi[\gamma_t \in E^{(a)}] = 1 - m$. We decompose Π into

$$\Pi = (1-m)\,\Pi^{(a)} + m\,\Pi^{(s)},$$

where

$$\Pi^{(a)}(d\gamma) = \frac{1_{[\gamma_t \in E^{(a)}]}\Pi(d\gamma)}{\Pi[\gamma_t \in E^{(a)}]}, \qquad \Pi^{(s)}(d\gamma) = \frac{1_{[\gamma_t \in E^{(s)}]}\Pi(d\gamma)}{\Pi[\gamma_t \in E^{(s)}]}.$$

For any $s \in [0,1]$, let further

$$\mu_s^{(a)} = (e_s)_\# \Pi^{(a)}, \qquad \mu_s^{(s)} = (e_s)_\# \Pi^{(s)},$$

and similarly

$$\pi^{(a)} = (e_0, e_1)_\# \Pi^{(a)}, \qquad \pi^{(s)} = (e_0, e_1)_\# \Pi^{(s)}.$$

Since it has been obtained by conditioning of a dynamical optimal transference plan, $\Pi^{(a)}$ is itself a dynamical optimal transference plan, and by construction $\mu_t^{(a)}$ is the absolutely continuous part of μ_t, while $\mu_t^{(s)}$ is its singular part. So the result of Step 2 applies to the path $(\mu_s^{(a)})_{0 \le s \le 1}$:

$$U_\nu(\mu_t^{(a)}) \le (1-t)\, U^{\beta_{1-t}}_{\pi^{(a)},\nu}(\mu_0^{(a)}) \;+\; t\, U^{\beta_t}_{\check{\pi}^{(a)},\nu}(\mu_1^{(a)}).$$

Actually, we shall not apply this inequality with the nonlinearity U, but rather with $U_m(r) = U((1-m)r)$, which obviously lies in $\mathcal{DC}_N$ if U does. So

$$(U_m)_\nu(\mu_t^{(a)}) \le (1-t)\, (U_m)^{\beta_{1-t}}_{\pi^{(a)},\nu}(\mu_0^{(a)}) \;+\; t\, (U_m)^{\beta_t}_{\check{\pi}^{(a)},\nu}(\mu_1^{(a)}). \tag{30.45}$$

Since $\mu_t^{(s)}$ is purely singular and $\mu_t = (1-m)\,\mu_t^{(a)} + m\,\mu_t^{(s)}$, the definition of U_ν implies

$$U_\nu(\mu_t) = (U_m)_\nu(\mu_t^{(a)}) + m\,U'(\infty). \tag{30.46}$$

By Theorem 30.17(iii), also $\mu_0^{(s)}$ is purely singular. So $\mu_0 = (1-m)\,\mu_0^{(a)} + m\,\mu_0^{(a)}$ implies

$$U^{\beta_{1-t}}_{\pi,\nu}(\mu_0) = (U_m)^{\beta_{1-t}}_{\pi^{(a)},\nu}(\mu_0^{(a)}) \;+\; m\,U'(\infty). \tag{30.47}$$

Similarly,

$$U^{\beta_t}_{\check{\pi},\nu}(\mu_1) = (U_m)^{\beta_t}_{\check{\pi}^{(a)},\nu}(\mu_1^{(a)}) \;+\; m\,U'(\infty). \tag{30.48}$$

The combination of (30.45), (30.46), (30.47) and (30.48) implies

$$U_\nu(\mu_t) \le (1-t)\, U^{\beta_{1-t}}_{\pi,\nu}(\mu_0) \;+\; U^{\beta_t}_{\check{\pi},\nu}(\mu_1).$$

This concludes the proof. □

Locality

Locality is one of the most fundamental properties that one may expect from any notion of curvature. In the setting of weak $\mathrm{CD}(K,N)$ spaces, the locality problem may be loosely formulated as follows: *If $(\mathcal{X},d,\nu)$ is weakly $\mathrm{CD}(K,N)$ in the neighborhood of any of its points, then $(\mathcal{X},d,\nu)$ should be a weakly $\mathrm{CD}(K,N)$ space.*

So far it is not known whether this "local-to-global" property holds true in general. However it is true at least in a *nonbranching space*, if either $K=0$ or $N=\infty$. The general case seems to depend on the following

Conjecture 30.27 (Local-to-global property in dimension 1). *Let $\theta\in(0,1)$ and $\alpha\in[0,\pi]$. Let $f:[0,1]\to\mathbb{R}_+$ be a measurable function such that for all $\lambda\in[0,1]$, $t,t'\in[0,1]$, the inequality*

$$f\big((1-\lambda)\,t+\lambda\,t'\big)\ge\ (1-\lambda)\,\left(\frac{\sin\big((1-\lambda)\,\alpha|t-t'|\big)}{(1-\lambda)\,\sin(\alpha|t-t'|)}\right)^{\theta}\,f(t)\ +\ \lambda\,\left(\frac{\sin\big(\lambda\,\alpha|t-t'|\big)}{\lambda\,\sin(\alpha|t-t'|)}\right)^{\theta}\,f(t') \tag{30.49}$$

holds true as soon as $|t-t'|$ is small enough. Then (30.49) automatically holds true for all $t,t'\in[0,1]$.

The same if the function sin *is replaced by* sinh *and α is allowed to vary in $\mathbb{R}_+$.*

I really don't have much to support this conjecture, except that it would imply a really nice (in my taste) result. It might be trivially false or trivially true, but I was unable to prove or disprove it. (If it would hold true only under additional regularity assumptions such as local integrability or continuity of f, this might be fine.)

To understand the relation of (30.49) to optimal transport, take $\theta=1-1/N$, $\alpha=\sqrt{|K|/(N-1)}\,d(\gamma_0,\gamma_1)$, $f(t)=\rho_t(\gamma_t)^{-1/N}$. Write $I_t(\gamma_0,\gamma_t,\gamma_1)$ for the inequality appearing in (30.40). Then Conjecture 30.27, if true, means that this inequality is local, in the sense that if $I_t(\gamma_{t_0},\gamma_{(1-t)\,t_0+t\,t_1},\gamma_{t_1})$ holds true for $|t_0-t_1|$ small enough, then it holds true for all t_0,t_1, and in particular $t_0=0$, $t_1=1$.

There are at least two limit cases in which the statement of Conjecture 30.27 becomes true. The first one is for $\alpha=0$ and θ fixed (this corresponds to $\mathrm{CD}(0,N)$, $N=1/(1-\theta)$); the second one is the limit when $\theta\to1$, $\alpha\to0$ in such a way that $\alpha^2/(1-\theta)$ converges to a finite limit (this corresponds to $\mathrm{CD}(K,\infty)$, and the limit of $\alpha^2/(1-\theta)$ would be $K\,d(\gamma_0,\gamma_1)^2$). In the first case, Conjecture 30.49 reduces to the locality of the property of concavity:

$$f\big((1-\lambda)\,t+\lambda\,t'\big)\ge\ (1-\lambda)\,f(t)+\lambda\,f(t');$$

while in the second case, it reduces to the locality of the more general property of κ-concavity ($\kappa\in\mathbb{R}$):

$$f\big((1-\lambda)\,t+\lambda\,t'\big)\ge\ (1-\lambda)\,f(t)+\lambda\,f(t')+\frac{\kappa\,\lambda(1-\lambda)}{2}\,|t-t'|^2. \tag{30.50}$$

These properties do satisfy a local-to-global principle, for instance because they are equivalent to the differential inequality $f''\le0$, or $f''\le-\kappa$, to be understood in distributional sense.

To summarize: If $K=0$ (resp. $N=\infty$), inequality (30.40) (resp. (30.41)) satisfies a local-to-global principle; in the other cases we don't know.

Next, I shall give a precise definition of what it means to satisfy $\mathrm{CD}(K,N)$ locally:

Definition 30.28 (local $\mathrm{CD}(K,N)$ space). *Let $K \in \mathbb{R}$ and $N \in [1,\infty]$. A locally compact, Polish length space $(\mathcal{X},d)$ equipped with a locally finite measure ν is said to be a locally weak $\mathrm{CD}(K,N)$ space if for any $x_0 \in \mathcal{X}$ there is $r>0$ such that whenever μ_0, μ_1 are two absolutely continuous probability measures supported in $B_r(x_0)$, there is a displacement interpolation $(\mu_t)_{0\le t\le 1}$ joining μ_0 to μ_1, and an associated optimal coupling π, in such a way that for all $t\in[0,1]$ and for all $U \in \mathcal{DC}_N$,*

$$U_\nu(\mu_t) \le (1-t)\, U^{\beta^{(K,N)}_{1-t}}_{\pi,\nu}(\mu_0) + t\, U^{\beta^{(K,N)}_{t}}_{\check\pi,\nu}(\mu_1). \tag{30.51}$$

Remark 30.29. In the previous definition, one could also have imposed that the whole path $(\mu_t)_{0\le t\le 1}$ is supported in $B_r(x_0)$. Both formulations are equivalent: Indeed, if μ_0 and μ_1 are supported in $B_{r/3}(x_0)$ then all measures μ_t are supported in $B_r(x_0)$.

Now comes the main result in this section:

Theorem 30.30 (from local to global $\mathrm{CD}(K,N)$). *Let $K\in\mathbb{R}$ and $N\in[1,\infty]$. If $K=0$ or $N=\infty$, any nonbranching locally weak $\mathrm{CD}(K,N)$ space is also a weak $\mathrm{CD}(K,N)$ space. The same is true for all values of K and N if Conjecture 30.27 has an affirmative answer.*

As in the proof of Theorem 30.26, one of the main ideas in the proof of Theorem 30.30 consists in using the nonbranching condition to translate integral conditions into pointwise density bounds along geodesic paths. Another idea consists in "cutting" dynamical optimal transference plans into small pieces, each of which is "small enough" that the local displacement convexity can be applied. The fact that we work along geodesic paths parametrized by $[0,1]$ explains that the whole locality problem is reduced to the one-dimensional "local-to-global" problem exposed in Conjecture 30.27.

Proof of Theorem 30.30. If we can treat the case $N>1$, then the case $N=1$ will follow by just letting N go to 1. (If $\mathcal{X}$ is $\mathrm{CD}(K,N)$ for all $N>1$ then it is also $\mathrm{CD}(K,1)$.) So I shall assume $N>1$; then by the nonbranching assumption, Theorem 30.26 applies. In the sequel, I shall use the shorthand $\beta_t = \beta^{(K,N)}_t$.

I shall also assume $N<\infty$; the case $N=\infty$ can be treated similarly, the main modification being the replacement of (30.40) by (30.41).

Let $(\mathcal{X},d,\nu)$ be a nonbranching locally weak $\mathrm{CD}(K,N)$ space. By repeating the proof of Theorem 30.26, we can show that for any $x_0\in\mathcal{X}$ there is $r=r(x_0)>0$ such that *any* displacement interpolation $(\mu_t)_{0\le t\le 1}$ which is supported in $B(x_0,r)$ satisfies (30.51). Moreover, if Π is a dynamical optimal transference plan such that $(e_t)_\#\Pi = \mu_t$, and each measure μ_t is absolutely continuous with density ρ_t, then $\Pi(d\gamma)$-almost all geodesics will satisfy inequality (30.40), which I recast below:

$$\frac{1}{\rho_t(\gamma_t)^{\frac{1}{N}}} \ge (1-t)\left(\frac{\beta_{1-t}(\gamma_0,\gamma_1)}{\rho_0(\gamma_0)}\right)^{\frac{1}{N}} + t\left(\frac{\beta_t(\gamma_0,\gamma_1)}{\rho_1(\gamma_1)}\right)^{\frac{1}{N}}. \tag{30.52}$$

Let μ_0, μ_1 be two absolutely continuous, compactly supported probability measures on $\mathcal{X}$, and let $B=B(z,R)$ be a large ball such that any geodesic going from $\mathrm{Spt}\,\mu_0$ to $\mathrm{Spt}\,\mu_1$ lies within B. Let Π be a dynamical optimal transference plan between μ_0 and μ_1. The goal is to prove that for all $U\in\mathcal{DC}_N\cap \mathrm{Lip}(\mathbb{R}_+,\mathbb{R}_+)$,

$$U_\nu(\mu_t) \le (1-t)\, U^{\beta_{1-t}}_{\pi,\nu}(\mu_0) + t\, U^{\beta_t}_{\check\pi,\nu}(\mu_1). \tag{30.53}$$

Then the conclusion will follow from Proposition 29.10.

The plan is to cut Π into very small pieces, each of which will be included in a sufficiently small ball that the local weak $\mathrm{CD}(K,N)$ criterion can be used. I shall first proceed to construct these small pieces.

Cover $\overline{B}(z,R)$ by a finite number of balls $B(x_j, r_j/3)$ with $r_j = r(x_j)$, and let $\overline{r} := \inf(r_j/3)$. For any $y \in B(z,R)$, the ball $B(y,\overline{r})$ lies inside some $B(x_j, r_j)$; so if $(\overline{\mu}_t)_{0\le t\le 1}$ is any displacement interpolation supported in some ball $B(y,\overline{r})$, $\overline{\Pi}$ is an associated dynamical optimal transference plan, and $\overline{\mu}_0$, $\overline{\mu}_1$ are absolutely continuous, then the density $\overline{\rho}_t$ of $\overline{\mu}_t$ will satisfy the inequality

$$\frac{1}{\overline{\rho}_t(\gamma_t)^{\frac{1}{N}}} \ge (1-t)\left(\frac{\beta_{1-t}(\gamma_0,\gamma_1)}{\overline{\rho}_0(\gamma_0)}\right)^{\frac{1}{N}} + t\left(\frac{\beta_t(\gamma_0,\gamma_1)}{\overline{\rho}_1(\gamma_1)}\right)^{\frac{1}{N}}, \tag{30.54}$$

$\overline{\Pi}(d\gamma)$-almost surely. The problem now is to cut Π into many small subplans $\overline{\Pi}$ and to apply (30.54) to all these subplans.

Let $\delta \in 1/\mathbb{N}$ be small enough that $4R\,\delta \le \overline{r}/3$, and let $B(y_\ell,\delta)_{1\le\ell\le L}$ be a finite covering of $B(z,R)$ by balls of radius δ. We define $A_1 = B(y_1,\delta)$, $A_2 = B(y_2,\delta)\setminus A_1$, $A_3 = B(y_3,\delta)\setminus(A_1\cup A_2)$, etc. This provides a covering of $B(z,R)$ by *disjoint* sets $(A_\ell)_{1\le\ell\le L}$, each of which is included in a ball of radius δ. (Without loss of generality, we can assume that they are all nonempty.)

Let $m = 1/\delta \in \mathbb{N}$. We divide the set Γ of all geodesics going from $\mathrm{Spt}\,\mu_0$ to $\mathrm{Spt}\,\mu_1$ into pieces, as follows. For any finite sequence $\underline{\ell} = (\ell_0,\ell_1,\dots,\ell_m)$, let

$$\Gamma_{\underline{\ell}} = \Big\{\gamma\in\Gamma;\ \ \gamma_0\in A_{\ell_0},\ \gamma_\delta\in A_{\ell_1},\ \gamma_{2\delta}\in A_{\ell_2},\dots,\ \gamma_{m\delta}=\gamma_1\in A_{\ell_m}\Big\}.$$

The sets $\Gamma_{\underline{\ell}}$ are disjoint. We discard the sequences $\underline{\ell}$ such that $\Pi[\Gamma_{\underline{\ell}}] = 0$. Let then $Z_{\underline{\ell}} = \Pi[\Gamma_{\underline{\ell}}]$, and let

$$\Pi_{\underline{\ell}} = \frac{1_{\Gamma_{\underline{\ell}}}\,\Pi_{\underline{\ell}}}{Z_{\underline{\ell}}}$$

be the law of γ conditioned by the event $\{\gamma\in\Gamma_{\underline{\ell}}\}$. Let further $\mu_{\underline{\ell},t} = (e_t)_\#\Pi_{\underline{\ell}}$, and $\pi_{\underline{\ell}} = (e_0,e_1)_\#\Pi_{\underline{\ell}}$.

For each $\underline{\ell}$ and $k\in\{0,\dots,m-2\}$, we define $\Pi^k_{\underline{\ell}}$ to be the image of $\Pi_{\underline{\ell}}$ by the restriction map $[0,1]\to[k\delta,(k+2)\delta]$. Up to affine reparametrization of time, $\Pi^k_{\underline{\ell}}$ is a dynamical optimal transference plan between the measures $\mu_{\underline{\ell},k\delta}$ and $\mu_{\underline{\ell},(k+2)\delta}$ (Theorem 7.27(i)-(ii)).

Let γ be a random geodesic distributed according to the law $\Pi^k_{\underline{\ell}}$. Almost surely, $\gamma(k\delta)$ belongs to A_{ℓ_k}, which has diameter at most $\overline{r}/3$. Moreover, the speed of γ is bounded above by $\mathrm{diam}\,(B(z,R)) \le 2R$, so on the time-interval $[k\delta,(k+2)\delta]$, γ moves at most by a distance $(2\delta)(2R)\le\overline{r}/3$. So γ is entirely contained in a set of diameter $2\overline{r}/3$. In particular, $\mu^k_{\underline{\ell},t}$ is entirely supported in a set of diameter $\overline{r}$, and satisfies the displacement convexity inequalities which are typical of the curvature-dimension bound $\mathrm{CD}(K,N)$.

By Theorem 7.27(iii), $\mu^k_{\underline{\ell},t}$ is (up to time-reparametrization) the unique optimal dynamical transference plan between $\mu_{\underline{\ell},k\delta}$ and $\mu_{\underline{\ell},(k+2)\delta}$. So by Theorem 30.17(ii), the absolute continuity of $\mu_{\underline{\ell},k\delta}$ implies the absolute continuity of $\mu_{\underline{\ell},t}$ for all $t\in[k\delta,(k+2)\delta)$. Since $\mu_{\underline{\ell},0}$ is absolutely continuous, an immediate induction shows that $\mu_{\underline{\ell},t}$ is absolutely continuous for all times. So we can apply (30.54) to each path $(\mu_{\underline{\ell},t})_{k\delta\le t\le(k+2)\delta}$; after time reparametrization, this becomes:

$$\forall k\in\{0,\dots,m-2\},\quad \Pi_{\underline{\ell}}(d\gamma)\text{-almost surely},\quad \forall t\in[0,1],\ \forall(t_0,t_1)\in[k\delta,(k+2)\delta],$$
$$\frac{1}{\rho_{\underline{\ell},(1-t)t_0+tt_1}\big(\gamma_{(1-t)t_0+tt_1}\big)^{\frac{1}{N}}} \ge (1-t)\Big(\frac{\beta_{1-t}(\gamma_{t_0},\gamma_{t_1})}{\rho_{\underline{\ell},t_0}(\gamma_{t_0})}\Big)^{\frac{1}{N}} + t\Big(\frac{\beta_t(\gamma_{t_0},\gamma_{t_1})}{\rho_{\underline{\ell},t_1}(\gamma_{t_1})}\Big)^{\frac{1}{N}}.$$

It follows that

$$\Pi_{\underline{\ell}}(d\gamma)\text{-almost surely}, \qquad \forall t \in [0,1], \quad \forall t_0, t_1 \in [0,1],$$
$$|t_0 - t_1| \le \delta \Longrightarrow \quad \frac{1}{\rho_{\underline{\ell},(1-t)t_0+tt_1}(\gamma_{(1-t)t_0+tt_1})^{\frac{1}{N}}} \ge (1-t)\left(\frac{\beta_{1-t}(\gamma_{t_0},\gamma_{t_1})}{\rho_{\underline{\ell},t_0}(\gamma_{t_0})}\right)^{\frac{1}{N}} + t\left(\frac{\beta_t(\gamma_{t_0},\gamma_{t_1})}{\rho_{\underline{\ell},t_1}(\gamma_{t_1})}\right)^{\frac{1}{N}}. \tag{30.55}$$

Inequality (30.55) is satisfied when t_0 and t_1 are close enough. Then our assumptions, and the discussion following Conjecture 30.27, imply that the same inequality is satisfied for all values of t_0 and t_1 in $[0,1]$. In particular, $\Pi_{\underline{\ell}}$-almost surely,

$$\frac{1}{\rho_{\underline{\ell},t}(\gamma_t)^{\frac{1}{N}}} \ge (1-t)\left(\frac{\beta_{1-t}(\gamma_0,\gamma_1)}{\rho_{\underline{\ell},0}(\gamma_0)}\right)^{\frac{1}{N}} + t\left(\frac{\beta_t(\gamma_0,\gamma_1)}{\rho_{\underline{\ell},1}(\gamma_1)}\right)^{\frac{1}{N}}. \tag{30.56}$$

By reasoning as in the proof of Theorem 30.26, we get the inequality

$$U_\nu(\mu_{\underline{\ell},t}) \le (1-t)\, U^{\beta_{1-t}}_{\pi_{\underline{\ell}},\nu}(\mu_{\underline{\ell},0}) + t\, U^{\beta_t}_{\check{\pi}_{\underline{\ell}},\nu}(\mu_{\underline{\ell},1}). \tag{30.57}$$

Recall that $\mu_t = \sum Z_{\underline{\ell}}\, \mu_{\underline{\ell},t}$; so the issue is now to add up the various contributions coming from different values of $\underline{\ell}$.

For each $\underline{\ell}$, we apply (30.57) with U replaced by $U_{\underline{\ell}} = U(Z_{\underline{\ell}}\cdot)/Z_{\underline{\ell}}$. Then, with the shorthand $U_{\underline{\ell},\nu} = (U_{\underline{\ell}})_\nu$ and $U^\beta_{\underline{\ell},\pi_{\underline{\ell}},\nu} = (U_{\underline{\ell}})_{\pi_{\underline{\ell}},\nu}$, we obtain

$$U_{\underline{\ell},\nu}(\mu_{\underline{\ell},t}) \le (1-t)\, U^{\beta_{1-t}}_{\underline{\ell},\pi_{\underline{\ell}},\nu}(\mu_{\underline{\ell},0}) + t\, U^{\beta_t}_{\underline{\ell},\check{\pi}_{\underline{\ell}},\nu}(\mu_{\underline{\ell},1}). \tag{30.58}$$

For any $t \in (0,1)$, the map $\gamma_t \to \gamma$ is injective, as a consequence of Theorem 7.27(iv), and in particular the measures $\mu_{\underline{\ell},t}$ are mutually singular as $\underline{\ell}$ varies. Then it follows from Lemma 29.5 that

$$U_\nu(\mu_t) = \sum_{\underline{\ell}} Z_{\underline{\ell}}\, U_{\ell,\nu}(\mu_{\underline{\ell},t}). \tag{30.59}$$

Since $\pi = \sum_{\underline{\ell}} Z_{\underline{\ell}}\, \pi_{\underline{\ell}}$, Lemma 29.5 also implies

$$\sum_{\underline{\ell}} U^{\beta_{1-t}}_{\underline{\ell},\pi_{\underline{\ell}},\nu}(\mu_{\underline{\ell},0}) = U^{\beta_{1-t}}_{\pi,\nu}(\mu_0);$$
$$\sum_{\underline{\ell}} U^{\beta_t}_{\underline{\ell},\check{\pi}_{\underline{\ell}},\nu}(\mu_{\underline{\ell},1}) = U^{\beta_t}_{\check{\pi},\nu}(\mu_1). \tag{30.60}$$

The combination of (30.58), (30.59) and (30.60) implies the desired conclusion (30.53). □

Appendix: Regularization in metric-measure spaces

Truncation and regularization by convolution are basic tools in real analysis. They are still available, to some extent, in metric-measure spaces, as I shall explain in this appendix. Recall that a boundedly compact metric space is a metric space in which closed balls are compact; and a locally finite measure is a measure which gives finite mass to balls.

Definition 30.31 (Truncation operators). *Let $(\mathcal{X}, d)$ be a boundedly compact metric space, and let $\star$ be an arbitrary base point. For any $R > 0$, let B_R be the closed ball $B_{R]}(\star)$. A $\star$-cutoff is a family of nonnegative continuous functions $(\chi_R)_{R>0}$ such that $1_{B_R} \le \chi_R \le 1_{B_{R+1}}$ for all R.*

More explicitly: χ_R is valued in $[0,1]$, and χ_R is identically equal to 1 on B_R, identically equal to 0 on B_{R+1}.

It is a consequence of Urysohn's lemma that a $\star$-cutoff always exists.

If μ is any finite measure on $\mathcal{X}$, then $\chi_R\,\mu$ converges to μ in total variation norm; moreover, for any $R > 0$, the truncation operator $T_R : \mu \to \chi_R\,\mu$ is a (nonstrict) contraction. As a particular case, if ν is any measure on $\mathcal{X}$, and $f \in L^1(\mathcal{X}, \nu)$, then $\chi_R\, f$ converges to f in $L^1(\nu)$.

A consequence of these results is the density of $C_c(\mathcal{X})$ in $L^1(\mathcal{X}, \nu)$, as soon as ν is locally finite. Indeed, if f is given in $L^1(\mathcal{X}, \nu)$, first choose R such that $\|f\|_{L^1(\mathcal{X}\setminus B_R)} \le \delta$; then pick up $g \in C(B_{R+1})$ such that $\|f - g\|_{L^1(B_{R+1},\nu)} \le \delta$. (Since B_{R+1} is compact, this can be done with basically Lusin's theorem, as recalled in the bibliographical notes of the previous chapter.) Finally define $\widetilde{g} := g\,\chi_R$, extended by 0 outside of B_R: then $\widetilde{g}$ is a continuous function with compact support, and it is easy to check that $\|f - \widetilde{g}\|_{L^1(\mathcal{X})} \le 2\delta$.

Now comes a more subtle technical tool:

Definition 30.32 (Regularizing kernels). *Let $(\mathcal{X}, d)$ be a boundedly compact metric space equipped with a locally finite measure ν, and let $\mathcal{Y}$ be a compact subset of $\mathcal{X}$. A $(\mathcal{Y}, \nu)$-regularizing kernel is a family of nonnegative continuous symmetric functions $(K_\varepsilon)_{\varepsilon>0}$ on $\mathcal{X} \times \mathcal{X}$, such that*

(i) $\forall x \in \mathcal{Y}, \qquad \displaystyle\int_{\mathcal{X}} K_\varepsilon(x,y)\,\nu(dy) = 1;$

(ii) $d(x,y) > \varepsilon \Longrightarrow \quad K_\varepsilon(x,y) = 0.$

It can be proven that a $(\mathcal{Y}, \nu)$-regularizing kernel always exists if $\mathcal{Y} \subset \mathrm{Spt}\,\nu$. The recipe is roughly as follows: Cover $\mathcal{Y}$ by a finite number of balls $B(x_i, \varepsilon/2)$, introduce a subordinate partition of unity (ϕ_i), and let

$$K_\varepsilon(x,y) := \sum_i \frac{\phi_i(x)\,\phi_i(y)}{\int \phi_i\, d\nu}. \tag{30.61}$$

See the bibliographical notes for more details.

Then, as soon as μ is a finite measure on $\mathcal{X}$, one may define a *continuous* function $K_\varepsilon\mu$ on $\mathcal{X}$ by

$$(K_\varepsilon\mu)(x) := \int_{\mathcal{X}} K_\varepsilon(x,y)\,\mu(dy).$$

Further, if $f \in L^1(\mathcal{X}, \nu)$, define $K_\varepsilon f := K_\varepsilon(f\nu)$.

The linear operator $K_\varepsilon : \mu \to K_\varepsilon\mu$ is mass-preserving, in the sense that for any nonnegative finite measure μ on $\mathcal{Y}$, one has $((K_\varepsilon\mu)\nu)[\mathcal{Y}] = \mu[\mathcal{Y}]$. More generally, K_ε defines a (nonstrict) contraction operator on $M(\mathcal{Y})$. Moreover, as $\varepsilon \to 0$,

- If $f \in C(\mathcal{X})$, then $K_\varepsilon f$ converges *uniformly* to f on $\mathcal{Y}$;

- If μ is a finite measure supported in $\mathcal{Y}$, then $(K_\varepsilon\mu)\nu$ converges weakly (against $C_b(\mathcal{X})$) to μ (this basically follows from the previous property by a duality argument);

- If $f \in L^1(\mathcal{Y})$, then $K_\varepsilon f$ converges to f in $L^1(\mathcal{Y})$ (this follows from the density of $C(\mathcal{Y})$ in $L^1(\mathcal{Y}, \nu)$, the fact that $K_\varepsilon f$ converges uniformly to f if f is continuous, and the contraction property of K_ε). There is in fact a more precise statement: For any $f \in L^1(\mathcal{Y}, \nu)$,

$$\int_{\mathcal{Y}\times\mathcal{Y}} |f(x) - f(y)|\,\nu(dx)\,\nu(dy) \xrightarrow[\varepsilon\to 0]{} 0.$$

Remark 30.33. If the measure ν is (locally) doubling, then one can ask more of the kernel (K_ε), than just properties (i) and (ii) in Definition 30.32. By Vitali's covering lemma, one can make sure that the covering $(B(x_i, \varepsilon/2))$ is such that the balls $B(x_i, \varepsilon/10)$ are disjoint. If (ϕ_i) is a partition of unity associated to the covering $(B(x_i, \varepsilon/2))$, necessarily ϕ_i is identically 1 on $B(x_i, \varepsilon/10)$, so $\int \phi_i \, d\nu \geq \nu[B(x_i, \varepsilon/10)] \geq C\nu[B(x_i, \varepsilon)]$, where C is a constant depending on the doubling constant of ν. The following uniform bound follows easily:

(iii) $K_\varepsilon(x, y) \leq \dfrac{C}{\nu[B_\varepsilon(x)]}$.

(Here C is another numerical constant, depending on the doubling constant of ν.) Assumptions (i), (ii) and (iii), together with the doubling property of ν, and classical Lebesgue density theory, guarantee that for any $f \in L^1(\mathcal{Y})$ the convergence of $K_\varepsilon f$ to f holds not only in $L^1(\mathcal{Y})$ but also *almost everywhere*.

Bibliographical Notes

Most of the material in this chapter comes from works by Lott and myself [247, 249, 248] and by Sturm [336, 340]. Some of the results are new.

Theorem 30.2 is taken from work by Lott and myself [247], as well as Theorems 30.19, 30.20, Corollary 30.11, and the first part of Theorem 30.24. Theorems 30.13, 30.10, 30.12 and 30.16 are due to Sturm [340, 336]. Part (i) of Theorem 30.17 was proven by Lott and myself in the case $K = 0$. Part (ii) follows a scheme of proof communicated to me by Sturm. Part (iv) is a well-known observation in $\mathbb{R}^n$, used in several recent works about optimal transport.

The second part of Theorem 30.24 is due to Lott and myself, it uses the extension of Propositttion 22.13 to metric-measure spaces [**?**]. It is shown there that the conclusions of Proposition 22.13 remain all true if $(\mathcal{X}, d)$ is a finite-dimensional Alexandrov space with curvature bounded below (this is a *pointwise* result, without any measure). On the other hand, if a measure ν on $\mathcal{X}$ satisfies (locally in $\mathcal{X}$) doubling and local Poincaré inequalities, then one can apply the technology developed by Cheeger [109] and others, to establish (22.51) ν-almost everywhere. (Also $|\nabla^- u| = |\nabla u|$ ν-almost everywhere if u is Lipschitz, so the distinction between $|\nabla u|$ and $|\nabla^- u|$ becomes irrelevant.)

The Poincaré inequalities appearing in Theorems 30.21 and 30.22 (in the case $K = 0$) are due to Lott and myself [249]. There the local Poincaré inequalities are proven in the formalism of upper gradients; this concept was put forward, among others, by Heinonen and Koskela [209], and played a key role in Cheeger's construction [109] of a differentiable structure on metric spaces satisfying a doubling condition and a local Poincaré inequality. Independently of [249], there were several simultaneous treatments of local Poincaré inequalities under $\mathrm{CD}(K, N)$ conditions, by Sturm [340] on one hand, and von Renesse [368] on the other. The proofs in all these works have many common points.

Inequality (30.29) was proven by Lott and myself in [247], at a time when we did not have the general definition of weak $\mathrm{CD}(K, N)$ spaces. The argument is inspired by previous works of Otto and myself [292, Theorem 4], and Ledoux [235]. It might still have some interest, since there is no reason why $\mathrm{CD}(0, N)$ and $\mathrm{CD}(K, \infty)$ together would should imply $\mathrm{CD}(K, N)$.

The second part of the conclusion of Theorem 30.8 is taken as the definition of the weak $\mathrm{CD}(K, \infty)$ criterion in [247], and the first part is taken as the definition of the general weak

$\mathrm{CD}(K,N)$ criterion in [249]. So Theorem 30.8 proves that the definition of weak $\mathrm{CD}(K,N)$ space used in these notes is equivalent to the definition used in [249].

The scheme of proof of Theorem 30.8 uses the approximation of general nonlinearities $U \in \mathcal{DC}_N$ by nonnegative Lipschitz continuous nonlinearities, which gives a quite simple proof of Theorem 30.6. This strategy however does not answer the natural question whether the conclusions of Theorem 30.6 still hold true for more general nonlinearities.

For Part (i) of Theorem 30.6, the Lipschitz continuity assumption can indeed be dispended with, although one should still impose some restrictions on the behavior of U close to 0, and some moment condition on μ, as in Theorem 30.4; this is worked out in [247, Appendix E].

For Part (ii) of Theorem 30.6, the answer is also affirmative for a constant distortion coefficient, as shown by the argument in [247, Appendix C]. The key idea is to replace the L^1 continuity argument by Jensen's inequality, in the form

$$U\left(\int_{\mathcal{X}} K_\varepsilon(x,y)\,\rho(y)\,\nu(dy)\right) \le \int_{\mathcal{X}} K_\varepsilon(x,y)\,U(\rho(y))\,\nu(dy);$$

after integration against ν this leads to

$$U_\nu(K_\varepsilon(\rho\nu)) \le U_\nu(\rho\nu).$$

(This argument uses the full strength of the regularizing kernels.)

When β is not constant things are much more subtle and after substantial effort I could solve the problem only under the restriction that U grows "at most polynomially" at infinity. When U is differentiable, this condition reads just

$$rU'(r) \le C[U_+(r) + r];$$

it is satisfied by any convex polynomial, but not by the exponential function. This assumption basically implies that $\beta\,U(\rho/\beta)$ has a locally Lipschitz dependence on β, with L^1 coefficients, which makes it possible to regularize the distortion coefficients into coefficients β_ε obtained by the action of the kernels K_ε. Then one can use Jensen's inequality, applied to the jointly convex function $(b,r) \to b\,U(r/b)$, to get, with obvious notation,

$$(K_\varepsilon\beta)\,U\left(\frac{K_\varepsilon\rho}{K_\varepsilon\beta}\right) \le K_\varepsilon\left(\beta\,U\left(\frac{\rho}{\beta}\right)\right).$$

Then a few additional technical steps are in order to take care of the dependence of $U^\beta_{\pi,\nu}$ on π. This approach yields more complete results at the level of Theorem 29.16(iii) and Theorem 30.6(ii), but it is much more complicated, which is why I abandoned it in the present set of notes.

Theorem 30.26 grew out of a joint work of mine with Figalli [**?**]; there we proved Theorem 30.26 (at least the parts which were not proven in [247]) in smooth Riemannian manifolds. The proof is slightly different from the one which I gave here; it uses Lemma 29.5.

In the case of Alexandrov spaces, the locality of the notion "curvature is bounded below by κ" was proven in full generality by Perelman [**?**]; see [**?**, ...].

The locality of $\mathrm{CD}(K,\infty)$ in nonbranching spaces was proven by Sturm [336], with a different argument than the one which I presented here. The rest of Theorem 30.30 is new.

When one restricts to $\lambda = 1/2$, Conjecture 30.27 takes a simpler form, and at least seems to be true for all θ outside $(0,1)$. But of course we are a priori only interested in the range $\theta \in (0,1)$.

I shall conclude with some technical notes.

The Dunford–Pettis theorem provides a sufficient condition for uniform equi-integrability: If a family $\mathcal{F} \subset L^1(\nu)$ is weakly sequentially compact in $L^1(\nu)$, then there exists a function $\Psi : \mathbb{R}_+ \to \mathbb{R}_+$ such that $\Psi(r)/r \to \infty$ and $\sup_{f\in\mathcal{F}} \int \Psi(f)\, d\nu < +\infty$. A proof can be found e.g. in [**?**], or in my own course on integration, available online at `www.umpa.ens-lyon.fr/~cvillani`.

Urysohn's lemma [147, Theorem 2.6.3] states the following: If $(\mathcal{X}, d)$ is a locally compact metric space (or even just a locally compact Hausdorff space), K is a compact subset of $\mathcal{X}$ and O is an open subset of $\mathcal{X}$ with $K \subset O$, then there is $f \in C_c(\mathcal{X})$ with $1_K \le f \le 1_O$. Although I did not need it in these notes, I also quote the Tietze-Urysohn extension theorem, which is often useful in the analysis of metric spaces: If $(\mathcal{X}, d)$ is a metric space, F is a closed subset of $\mathcal{X}$, and $f : F \to \mathbb{R}$ is uniformly continuous on F, then it is possible to extend f into a continuous function on the whole of $\mathcal{X}$, with preservation of the supremum norm of f; see [147, Theorem 2.6.4].

I learnt from Lott the nice recipe (30.61) to construct regularizing kernels. In [247, Appendix C], we worked out in detail the properties stated after Definition 30.32. We used this tool extensively in our work, and also discuss regularization in noncompact spaces. Even in the framework of absolutely continuous measures, the approach based on the regularizing kernel has many advantages over Lusin's approximation theorem (linearity, preservation of convexity inequalities, etc.).

Conclusions and open problems

In these notes I have tried to present a consistent picture of the theory of optimal transport, with a *dynamical*, *probabilistic* and *geometric* point of view, insisting on the concepts of displacement interpolation, probabilistic representation, and curvature effects.

The qualitative description of optimal transport, developed in Part I, now seems to be relatively well understood, but only at the price of (a) working in a reasonably smooth ambiant space; (b) forgetting about the regularity of optimal transport. Of course one cannot eliminate both restrictions at the same time, but it would be desirable to understand how to relax either one of them. Possible directions of research include:

(a) Establishing representation theorems for the optimal transport in Alexandrov spaces, or approximate representation theorems for the optimal transport in more singular spaces. A related problem is how far one can push the machinery of changes of variables and curvature bounds which was used in Chapter 17 to establish displacement convexity theorems. A preliminary step in that direction might be nonsmooth analogues of Mather's shortening lemma, as stated for instance in Open Problem 8.21.

(b) Establishing smoothness theorems for the optimal transport on smooth manifolds, under adequate structure conditions. On this last topic, a lot of progress has been made recently by Neil Trudinger and his collaborators on one hand, and Grégoire Loeper on the other hand. In this business, Assumption (C) in Chapter 9, and the complicated fourth-order differential conditions found by Trudinger and collaborators, seem to play a crucial role. At the time of writing, Loeper seems to be in possession of an argument showing that the Euclidean sphere does satisfy this mysterious condition, giving hope for further developments. In the present book, regularity theory was not needed, but its understanding seems compulsory for various applications, such as the analysis of continuous numerical schemes, as studied recently for instance by Francesca Rapetti and Grégoire Loeper.

In this discussion I am implicitly assuming that the cost function satisfies some kind of "strict convexity" property; for instance that it is associated with a Lagrangian that is strictly convex in the velocity variable, as in Chapters 8 to 10. But there are important cost functions which do not satisfy this assumption at all, such as the plain distance function. The structure of the optimal transport for such cost functions has received a lot of attention, with some important recent progress in connection with the Aubry-Mather theory; see the bibliographical notes of Chapter 10 for more information. Actually, a whole monograph could be written on that problem by now.

As for the applications of optimal transport to Riemannian geometry, a consistent picture is also emerging, as I have tried to show in Part II. The main regularity problems seem to be under control here, but there remain several challenging "structural" problems:

- How can one best understand the relation between plain displacement convexity and distorted displacement convexity, as exposed in Chapter 17? Is there an Eulerian counterpart of the latter concept? See Open Problems 17.29 and 17.30 for more precise formulations.

- Optimal transport seems to work well to establish sharp geometric inequalities when the "natural dimension of the inequality" coincides with the dimension bound; on the other hand, so far it has failed to establish for instance sharp logarithmic Sobolev or Talagrand inequalities (infinite-dimensional) under a $\mathrm{CD}(K, N)$ curvature–dimension bound for $N < \infty$ (Open Problems 21.6 and 22.41). The sharp L^2-Sobolev inequality (21.10) has also escaped investigations based on optimal transport (Open Problem 21.10). Can one find a more precise strategy to attack such problems by a displacement convexity approach? A seemingly closely related question is whether one can mimick (maybe by changes of

unknowns in the transport problem??) the changes of variables in the Γ_2 formalism, which are often at the basis of the derivation of such sharp inequalities (as in e.g. the recent works of Jérôme Demange). To add to the confusion, there is the mysterious structure condition (25.9) found by Demange, and it is natural to ask whether this condition has any interpretation in terms of optimal transport.

- Are there interesting **examples of displacement convex functionals** apart from the ones that have already been explored during the past ten years, that are all basically of the form $\int_M U(\rho)\, d\nu + \int_{M^k} V\, d\mu^{\otimes k}$? It is frustrating that so few examples of displacement convex functionals are known, in contrast with the enormous amount of plainly convex functionals that one can construct. Open Problem 15.11 might be related to this question.

- Is there a transport-based proof of the **Lévy-Gromov isoperimetric inequalities** (Open Problem 21.15), that would not involve such a "hard analysis" as in the currently known arguments? Besides its intrinsic interest, such a proof could hopefully be adapted to nonsmooth spaces such as the weak $\mathrm{CD}(K, N)$ spaces studied in Part III.

- Caffarelli's **log concave perturbation theorem** (alluded to in Chapter 2) can be seen as another riddle in the picture. The Gaussian space can be seen as the infinite-dimensional version of the sphere, which is the Riemannian "reference space" with positive constant (sectional) curvature; and the space $\mathbb{R}^n$ equipped with a log concave measure is a space of nonnegative Ricci curvature. So Caffarelli's theorem can be formally restated as follows: *If the Euclidean space $(\mathbb{R}^n, d_2)$ is equipped with a probability measure ν that makes it a $\mathrm{CD}(K, \infty)$ space, then ν can be realized as a 1-Lipschitz push-forward of the reference Gaussian measure with curvature K.* This implies almost obviously that isoperimetric inequalities in $(\mathbb{R}^n, d_2, \nu)$ are not worse than isoperimetric inequalities in the Gaussian space; so there is a strong analogy between Caffarelli's theorem on one hand, and the Lévy-Gromov isoperimetric inequality on the other hand. It is natural to ask whether there is a common framework for both results; this does not seem obvious at all, and I have not been able to formulate even a decent guess of what could be a geometric generalization of Caffarelli's theorem.

- Another important remark is that the geometric theory has been almost exclusively developed in the case of the optimal transport with *quadratic* cost function; the exponent $p = 2$ here is natural in the context of Riemannian geometry, but working with other exponents might lead to new geometric territories. A related question is Open Problem 15.12.

In Part III of these notes, I discussed the emerging theory of weak Ricci curvature lower bounds in metric-measure spaces, based on displacement convexity inequalities. The theory has grown very fast and it is starting to be rather well-developed, however some challenging issues remain to be solved before one can consider it as mature. Here are three missing pieces of the puzzle:

- A **globalization theorem** that would play the role of the Toponogov-Perelman theorem for Alexandrov spaces with a lower bound on the curvature. This theorem should state essentially that if $(\mathcal{X}, d, \nu)$ is locally a weak $\mathrm{CD}(K, N)$ space, then it is globally a weak $\mathrm{CD}(K, N)$ space.

- The **compatibility with the theory of Alexandrov spaces**. Alexandrov spaces have proven their flexibility and have gained a lot of popularity among geometers. Since Alexandrov bounds are weak sectional curvature bounds, they should in principle be able to control weak Ricci curvature bounds. The natural question here can be stated as follows: Let $(\mathcal{X}, d)$ be a finite-dimensional Alexandrov space with dimension n and curvature

bounded below by κ, and let $\mathcal{H}^n$ be the n-dimensional Hausdorff measure; does it follow that $(\mathcal{X}, d, \mathcal{H}^n)$ is a weak $\mathrm{CD}((n-1)\kappa, n)$ space?

- A thorough discussion of the **branching** problem: Find examples of weak $\mathrm{CD}(K, N)$ spaces that are branching; that are singular but nonbranching; identify simple regularity conditions that prevent branching; etc.

It was realized in a recent Oberwolfach meeting, as a consequence of discussions between Dario Cordero-Erausquin, Karl-Theodor Sturm and myself, that the Euclidean space $\mathbb{R}^n$, equipped with any norm $\|\cdot\|$, is a weak $\mathrm{CD}(0, n)$ space:

Theorem. *Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$ (considered as a distance on $\mathbb{R}^n \times \mathbb{R}^n$), and let λ_n be the n-dimensional Lebesgue measure. Then the metric-measure space $(\mathbb{R}^n, \|\cdot\|, \lambda_n)$ is a weak $\mathrm{CD}(0, n)$ space in the sense of Definition 29.6.*

I preferred not to include this theorem in the body of these notes, because it appeals to some results that have not yet been adapted to a genuinely geometric context, and which I therefore preferred not to discuss. I shall sketch the proof at the end of this text, but before I would like to explain why this result is at the same time motivating, and a bit shocking:

(a) As pointed out to me by John Lott, if $\|\cdot\|$ is not Euclidean, then the metric-measure space $(\mathbb{R}^n, \|\cdot\|, \lambda_n)$ cannot be realized as a limit of smooth Riemannian manifolds with a uniform $\mathrm{CD}(0, N)$ bound, because it fails to satisfy the *splitting principle*. (If a nonnegatively curved space admits a line, i.e. a geodesic parametrized by $\mathbb{R}$, then the space can be "factorized" by this geodesic; results by Cheeger, Colding and Gromoll say that this is true for $\mathrm{CD}(0, N)$ manifolds and their measured Gromov-Hausdorff limits.)

(b) If $\|\cdot\|$ is not the Euclidean norm, the resulting metric space is very singular in certain respects: It is in general not an Alexandrov space, and it can be extremely branching. For instance, if $\|\cdot\|$ is the ℓ_∞ norm, then any two distinct points are joined by a non-countable infinity of geodesics. Since $(\mathbb{R}^n, \|\cdot\|_{\ell_\infty}, \lambda_n)$ is the (pointed) limit of the nonbranching spaces $(\mathbb{R}^n, \|\cdot\|_{\ell_p}, \lambda_n)$ as $p \to \infty$, we also realize that weak $\mathrm{CD}(K, N)$ bounds do not prevent the appearance of branching in measured Gromov-Hausdorff limits.

On the other hand, the study of optimal Sobolev inequalities in $\mathbb{R}^n$ which I performed together with Bruno Nazaret and Dario Cordero-Erausquin shows that optimal Sobolev inequalities basically do not depend on the choice of the norm on $\mathbb{R}^n$. In a Riemannian context, Sobolev inequalities strongly depend on Ricci curvature bounds; so, our result seems to indicate that it is not absurd after all to decide that $\mathbb{R}^n$ is a weak $\mathrm{CD}(0, n)$ space independently of the norm.

One can also ask whether there are additional regularity conditions that might be added to the definition of weak $\mathrm{CD}(K, N)$ space, in order to enforce nonbranching, or the splitting principle, or both, and in particular rule out non-Euclidean norms.

As a side consequence of point (a) above, we realize that smooth $\mathrm{CD}(K, N)$ manifolds are *not* dense in the spaces $\mathcal{CDD}(K, N, D, m, M)$ introduced in Theorem 29.23.

The interpretation of dissipative equations as gradient flows with respect to optimal transport, and the theory reviewed in Chapters 23 to 25, also leads to fascinating issues that are relevant in smooth or nonsmooth geometry as well as in partial differential equations. For instance,

(a) Can one define a reasonably well-behaved **heat flow** on weak $\mathrm{CD}(K, N)$ spaces by taking the gradient flow for Boltzmann's H functional? The theory of gradient flows in abstract metric spaces has been pushed very far, in particular with the above-mentioned

work of Ambrosio, Gigli and Savaré; so it might not be so difficult to define an object that would play the role of a heat semigroup; but this will be of limited value unless one can prove relevant theorems about it.

This problem might be related to the possibility of defining a Laplace operator on a singular space, an issue which has been addressed in particular by Jeff Cheeger and Toby Colding, for limits of Riemannian manifolds. However, their construction is strongly based on regularity properties enjoyed by such limits, and breaks down e.g. for $\mathbb{R}^n$ equipped with a non-Euclidean norm. So it might be hopeless to define a decent Laplace operator on general $\mathrm{CD}(K, N)$ spaces without any additional regularity structure.

(b) Can one extend the theory of dissipative equations to other equations, which are of Hamiltonian, or, even more interestingly, of dissipative Hamiltonian nature? As explained in the bibliographical notes of Chapter 23, there has been some recent work in that direction by Luigi Ambrosio, Wilfrid Gangbo and others, however the situation is still far from clear.

A loosely related issue is the study of the semi-geostrophic system, which in the simplest situations can formally be written as a Hamiltonian flow, where the Hamiltonian function is the square Wasserstein distance with respect to some uniform reference state. I think that the rigorous qualitative understanding of the semi-geostrophic system is one of the most exciting problems that I am aware of in theoretical fluid mechanics; and discussions with Mike Cullen convinced me that it is extremely relevant in applications to meteorology. Although the theory of the semi-geostrophic system is still full of fundamental open problems, enough has already been written on it to make the substance of a complete monograph.

On a much more theoretical level, the geometric understanding of the Wasserstein space $P_2(\mathcal{X})$, where $\mathcal{X}$ is a Riemannian manifold or just a length space, has been the object of several studies recently, and still retains many mysteries. For instance, there is a neat statement according to which $P_2(\mathcal{X})$ is nonnegatively curved, in the Alexandrov sense, if and only if $\mathcal{X}$ is itself nonnegatively curved. But there is no similar statement for nonzero lower bounds on the curvature! In fact, if x is a point of negative curvature, then the curvature of $P_2(\mathcal{X})$ seems to be unbounded in both directions ($+\infty$ and $-\infty$) in the neighborhood of δ_x. Also it is not clear what exactly is "the right" structure on, say, $P_2(\mathbb{R}^n)$ and recent works on the subject have provided differing answers. In their recent book about gradient flows, Luigi Ambrosio, Nicola Gigli and Giuseppe Savaré make an intriguing observation: It is possible to define "generalized geodesics" in $P_2(\mathbb{R}^n)$ by considering the law of $(1-t)X_0 + tX_1$, where (X_0, Z) and (X_1, Z) are optimal couplings. These generalized geodesics have intriguing properties: For instance, they still satisfy the characteristic displacement interpolation inequalities; and they provide curves of "nonpositive curvature", that can be exploited for various purposes, such as error estimates for approximate gradient flow schemes. I have no idea whether similar objects can be introduced in a genuinely geometric setting.

This list only provides a small sample of the many problems that remain open in the theory of optimal transport. I mainly focused on the problems that seemed most fundamental to me, knowingly forgetting many issues of interest, for instance those related to the theory of concentration of measure, or other applications to probability theory. At the time of writing, there is a lot of activity in that direction; as an example, I worked recently, together with François Bolley and Arnaud Guillin, on quantitative Sanov-type bounds for particle systems, with an approach based on optimal transport.

Another fundamental issue which I did not address at all is the subject of numerical analysis of optimal transport. This topic also has a long and complex history, with some famous schemes such as the old simplex algorithm, or the more recent auction algorithm. Recent works by Uriel Frisch and collaborators in cosmology provide an example where one would like to efficiently solve the optimal transport problem with really huge sets of data. To add to the variety of methods, continuous schemes based on partial differential equations have been making their way lately. All in all, this subject certainly deserves a reference book on its own, with experiments and systematic comparisons of algorithms, benchmark problems and so forth. By the way, the optimum matching problem is one of the subjects that Donald Knuth has planned to address in his long awaited Volume 4 of *The Art of Computer Programming*.

It is likely that many other unexpected developments will arise in the area of optimal transport. Actually, at the time when I was completing my first book on the subject, I certainly did not expect that it would continue to grow even much faster than before, and in so many directions.

Sketch of proof of the Theorem. First consider the case when $N = \|\cdot\|$ is a uniformly convex, smooth norm, in the sense that

$$\lambda\, I_n \le \nabla^2 N^2 \le \Lambda\, I_n$$

for some positive constants λ and Λ. Then the cost function $c(x,y) = N(x-y)^2$ is both strictly convex and $C^{1,1}$, i.e. uniformly semi-concave. This makes it possible to apply Theorem 10.26 (recall Example 10.30) and deduce the following theorem about the structure of optimal maps: Namely, if μ_0 and μ_1 are compactly supported and absolutely continuous, then there is a unique optimal map, and it takes the form

$$T(x) = x - \nabla (N^2)^*(-\nabla\psi(x)), \qquad \psi \text{ a } c\text{-convex function.}$$

Since the norm is uniformly convex, it is not branching, and geodesic lines are just straight lines; so the displacement interpolation takes the form $(T_t)_\#(\rho_0\, \lambda_n)$, where

$$T_t(x) = x - t\, \nabla (N^2)^*(-\nabla\psi(x)) \qquad 0 \le t \le 1.$$

Let $\theta(x) = \nabla (N^2)^*(-\nabla\psi(x))$. By [365, Remark 2.56], the Jacobian matrix $\nabla\theta$, although not symmetric, is pointwise diagonalizable, with eigenvalues bounded above by 1 (this remark goes back at least to a 1996 preprint by Otto [287]). Then it is easy to show that $t \to \det(I_n - t\nabla\theta)^{1/n}$ is a concave function of t [365, Lemma 5.21], and one can reproduce McCann's proof of displacement convexity for U_{λ_n}, as soon as $U \in \mathcal{DC}_n$ [365, Theorem 5.15 (i)].

This shows that $(\mathbb{R}^n, N, \lambda_n)$ satisfies the $\mathrm{CD}(0,n)$ displacement convexity inequalities when N is a smooth uniformly convex norm. Now if N is arbitrary, it can be approximated by a sequence $(N_k)_{k\in\mathbb{N}}$ of such norms, in such a way that $(\mathbb{R}^n, N, \lambda_n, 0)$ is the pointed measured Gromov-Hausdorff limit of $(\mathbb{R}^n, N_k, \lambda_n, 0)$ as $k \to \infty$. Then the general conclusion follows by stability of the weak $\mathrm{CD}(0,n)$ criterion. □

Remark. In the above argument the spaces $(\mathbb{R}^n, N_k, \lambda_n)$ satisfy the property that the displacement interpolation between any two absolutely continuous, compactly supported probability measures is unique; while the limit space $(\mathbb{R}^n, N, \lambda_n)$ does not necessarily satisfy this property. For instance, if $N = \|\cdot\|_{\ell_\infty}$, then there is an enormous amount of displacement interpolations between two given probability measures; and most of them do

not satisfy the displacement convexity inequalities that are used to define $\mathrm{CD}(0, n)$ bounds. This shows that if in Definition 29.6 one requires the inequality (29.7) to hold true for *any* Wasserstein geodesic, rather than for *some* Wasserstein geodesic, then the resulting $\mathrm{CD}(K, N)$ property is not stable under measured Gromov-Hausdorff convergence.

References

1. Agueh, M. *Existence of solutions to degenerate parabolic equations via the Monge-Kantorovich theory.* PhD thesis, Georgia Tech, Atlanta, 2001.
2. Agueh, M. Existence of solutions to degenerate parabolic equations via the Monge-Kantorovich theory. *Adv. Differential Equations 10*, 3 (2005), 309–360.
3. Agueh, M., Ghoussoub, N., and Kang, X. Geometric inequalities via a general comparison principle for interacting gases. *Geom. Funct. Anal. 14*, 1 (2004), 215–244.
4. Alberti, G., and Ambrosio, L. A geometrical approach to monotone functions in $\mathbb{R}^n$. *Math. Z. 230*, 2 (1999), 259–316.
5. Alberti, G., Ambrosio, L., and Cannarsa, P. On the singularities of convex functions. *Manuscripta Math. 76*, 3-4 (1992), 421–435.
6. Alesker, S., Dar, S., and Milman, V. A remarkable measure preserving diffeomorphism between two convex bodies in $\mathbf{r}^n$. *Geom. Dedicata 74*, 2 (1999), 201–212.
7. Alexandrov, A. Almost everywhere existence of the second differential of a convex function and some properties of convex surfaces connected to it (in Russian). *Uc̄ en. Zap. Leningrad. Gos. Univ. 37* (1939), 3–35.
8. Alexandrov, A. Existence and uniqueness of a convex surface with given integral curvature. *Dokl. Akad. Nauk. SSSR 35* (1942), 131–134.
9. Alexandrov, A. Smoothness of a convex surface of bounded gaussian curvature. *Dokl. Akad. Nauk. SSSR 36* (1942), 195–199.
10. Ambrosio, L. Lecture notes on optimal transport problems. Notes from a CIME summer school in Madeira, 2–9 July 2000. To appear in the Proceedings of the CIME Course held in Martina Franca (Italy), 2–9 September 2001, Lecture Notes in Mathematics.
11. Ambrosio, L. Steepest descent flows and applications to spaces of probability measures. Lecture notes for a Summer School in Santander, July 2004 (preliminary version).
12. Ambrosio, L. Transport equation and Cauchy problem for BV vector fields. *Invent. Math. 158*, 2 (2004), 227–260.
13. Ambrosio, L., Fusco, N., and Pallara, D. *Functions of bounded variation and free discontinuity problems.* Oxford Mathematical Monographs. The Clarendon Press Oxford University Press, New York, 2000.
14. Ambrosio, L., and Gangbo, W. Hamiltonian flows in $P_2(\mathbb{R}^{2d})$. Report at the MSRI Optimal Transport meeting, Berkeley, November 2005.
15. Ambrosio, L., Gigli, N., and Savaré, G. *Gradient flows in metric spaces and in the space of probability measures.* Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 2005.
16. Ambrosio, L., Kirchheim, B., and Pratelli, A. Existence of optimal transport maps for crystalline norms. *Duke Math. J. 125*, 2 (2004), 207–241.
17. Ambrosio, L., and Pratelli, A. Existence and stability results in the L^1 theory of optimal transportation. In *Optimal transportation and applications (Martina Franca, 2001)*, vol. 1813 of *Lecture Notes in Math.* Springer, Berlin, 2003, pp. 123–160.
18. Ambrosio, L., and Tilli, P. *Topics on analysis in metric spaces*, vol. 25 of *Oxford Lecture Series in Mathematics and its Applications.* Oxford University Press, Oxford, 2004.
19. Andreu, F., Caselles, V., Mazon, J., and Moll, S. Finite propagation speed for limited flux diffusion equations. To appear in *Arch. Rational Mech. Anal.*
20. Andreu, F., Caselles, V., and Mazón, J. M. The Cauchy problem for a strongly degenerate quasilinear equation. *J. Eur. Math. Soc. (JEMS) 7*, 3 (2005), 361–393.
21. Ané, C., Blachère, S., Chafaï, D., Fougères, P., Gentil, I., Malrieu, F., Roberto, C., and Scheffer, G. *Sur les inégalités de Sobolev logarithmiques*, vol. 10 of *Panoramas et Synthèses.* Société Mathématique de France, 2000.

22. Appell, P. Mémoire sur les déblais et les remblais des systèmes continus ou discontinus. *Mémoires présentés par divers Savants à l'Académie des Sciences de l'Institut de France, Paris 29* (1887), 1–208. Available online at `http:gallica.bnf.fr`.
23. Arnold, A., Markowich, P., Toscani, G., and Unterreiter, A. On logarithmic Sobolev inequalities and the rate of convergence to equilibrium for Fokker-Planck type equations. *Comm. Partial Differential Equations 26*, 1–2 (2001), 43–100.
24. Arnol′d, V. I. *Mathematical methods of classical mechanics*, vol. 60 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 199? Translated from the 1974 Russian original by K. Vogtmann and A. Weinstein, Corrected reprint of the second (1989) edition.
25. Aronson, D. G., and Bénilan, P. Régularité des solutions de l'équation des milieux poreux dans $\mathbf{R}^N$. *C. R. Acad. Sci. Paris Sér. A-B 288*, 2 (1979), A103–A105.
26. Aronsson, G. A mathematical model in sand mechanics: Presentation and analysis. *SIAM J. Appl. Math. 22* (1972), 437–458.
27. Aronsson, G., and Evans, L. C. An asymptotic model for compression molding. *Indiana Univ. Math. J. 51*, 1 (2002), 1–36.
28. Aronsson, G., Evans, L. C., and Wu, Y. Fast/slow diffusion and growing sandpiles. *J. Differential Equations 131*, 2 (1996), 304–335.
29. Attouch, L., and Soubeyran, A. From procedural rationality to routines: A "Worthwile to Move" approach of satisficing with not too much sacrificing. Preprint, 2005.
30. Bakry, D. L'hypercontractivité et son utilisation en théorie des semigroupes. In *Ecole d'été de Probabilités de Saint-Flour*, no. 1581 in Lecture Notes in Math. Springer, 1994.
31. Bakry, D., and Emery, M. Diffusions hypercontractives. In *Sém. Proba. XIX*, no. 1123 in Lecture Notes in Math. Springer, 1985, pp. 177–206.
32. Bakry, D., and Ledoux, M. A logarithmic Sobolev form of the Li-Yau parabolic inequalities. To appear in *Rev. Matematica Iberoamericana*; available online via `www.lsp.ups-tlse.fr/Ledoux/`.
33. Bakry, D., and Ledoux, M. Lévy-Gromov's isoperimetric inequality for an infinite-dimensional diffusion generator. *Invent. Math. 123*, 2 (1996), 259–281.
34. Bakry, D., and Qian, Z. Some new results on eigenvectors via dimension, diameter, and Ricci curvature. *Adv. Math. 155*, 1 (2000), 98–153.
35. Bakry, D., and Qian, Z. M. Harnack inequalities on a manifold with positive or negative Ricci curvature. *Rev. Mat. Iberoamericana 15*, 1 (1999), 143–179.
36. Bangert, V. Minimal measures and minimizing closed normal one-currents. *Geom. Funct. Anal. 9*, 3 (1999), 413–427.
37. Barles, G. *Solutions de viscosité des équations de Hamilton-Jacobi*, vol. 17 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Paris, 1994.
38. Benaïm, M., Ledoux, M., and Raimond, O. Self-interacting diffusions. *Probab. Theory Related Fields 122*, 1 (2002), 1–41.
39. Benaïm, M., and Raimond, O. Self-interacting diffusions. II. Convergence in law. *Ann. Inst. H. Poincaré Probab. Statist. 39*, 6 (2003), 1043–1055.
40. Benaïm, M., and Raimond, O. Self-interacting diffusions. II. Convergence in law. *Ann. Inst. H. Poincaré Probab. Statist. 39*, 6 (2003), 1043–1055.
41. Benaïm, M., and Raimond, O. Self-interacting diffusions. III. Symmetric interactions. *Ann. Probab. 33*, 5 (2005), 1717–1759.
42. Benamou, J.-D., and Brenier, Y. A numerical method for the optimal time-continuous mass transport problem and related problems. In *Monge Ampère equation: applications to geometry and optimization (Deerfield Beach, FL, 1997)*. Amer. Math. Soc., Providence, RI, 1999, pp. 1–11.
43. Benedetto, D., Caglioti, E., Carrillo, J. A., and Pulvirenti, M. A non-Maxwellian steady distribution for one-dimensional granular media. *J. Statist. Phys. 91*, 5-6 (1998), 979–990.
44. Benedetto, D., Caglioti, E., and Pulvirenti, M. A one dimensional Boltzmann equation with inelastic collisions. *Rend. Sem. Mat. Fis. Milano 67* (1997), 169–179 (2000).
45. Bénilan, P. Solutions intégrales d'équations d'évolution dans un espace de Banach. *C. R. Acad. Sci. Paris Sér. A-B 274* (1972), A47–A50.
46. Berkes, I., and Philipp, W. An almost sure invariance principle for the empirical distribution function of mixing random variables. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 41*, 2 (1977/78), 115–137.
47. Bernard, P., and Buffoni, B. Optimal mass transportation and Mather theory. Preprint. Available online at `http://www-fourier.ujf-grenoble.fr/~pbernard/publi.html`, 2004.
48. Bernard, P., and Buffoni, B. The Monge problem for supercritical Mañé potentials on compact manifolds. Preprint. Available online at `http://www-fourier.ujf-grenoble.fr/~pbernard/publi.html`, 2005.
49. Bernot, M., Caselles, V., and Morel, J.-M. Traffic plans. Preprint, 2004.

50. Bertsekas, D. *Network optimization: Continuous and discrete models*. Athena Scientific, 1998. Referenced online at `http://www.athenasc.com/netbook.html`.
51. Biler, P., Dolbeault, J., and Esteban, M. J. Intermediate asymptotics in L^1 for general nonlinear diffusion equations. *Appl. Math. Lett. 15*, 1 (2002), 101–107.
52. Billingsley, P. *Convergence of probability measures*, second ed. John Wiley & Sons Inc., New York, 1999. A Wiley-Interscience Publication.
53. Blower, G. The Gaussian isoperimetric inequality and transportation. *Positivity 7*, 3 (2003), 203–224.
54. Blower, G. Displacement convexity for the generalized orthogonal ensemble. *J. Statist. Phys. 116*, 5-6 (2004), 1359–1387.
55. Bobkov, S. A functional form of the isoperimetric inequality for the Gaussian measure. *J. Funct. Anal. 135*, 1 (1996), 39–49.
56. Bobkov, S., Gentil, I., and Ledoux, M. Hypercontractivity of Hamilton-Jacobi equations. *J. Math. Pures Appl. 80*, 7 (2001), 669–696.
57. Bobkov, S. G., and Götze, F. Exponential integrability and transportation cost related to logarithmic Sobolev inequalities. *J. Funct. Anal. 163*, 1 (1999), 1–28.
58. Bobkov, S. G., and Houdré, C. Some connections between isoperimetric and Sobolev-type inequalities. *Mem. Amer. Math. Soc. 129*, 616 (1997).
59. Bobkov, S. G., and Ledoux, M. From Brunn-Minkowski to Brascamp-Lieb and to logarithmic Sobolev inequalities. *Geom. Funct. Anal. 10*, 5 (2000), 1028–1052.
60. Bobylev, A., and Toscani, G. On the generalization of the Boltzmann H-theorem for a spatially homogeneous Maxwell gas. *J. Math. Phys. 33*, 7 (1992), 2578–2586.
61. Bolley, F. Separability and completeness for the Wasserstein distance, 2004. Personal communication.
62. Bolley, F., and Carrillo, J. Work in progress.
63. Bolley, F., and Villani, C. Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities. *Ann. Fac. Sci. Toulouse Math. (6) 14*, 3 (2005), 331–352.
64. Bonami, A. Étude des coefficients de Fourier des fonctions de $L^p(G)$. *Ann. Inst. Fourier (Grenoble) 20*, fasc. 2 (1970), 335–402 (1971).
65. Borell, C. Convex set functions in d-space. *Period. Math. Hungar. 6*, 2 (1975), 111–136.
66. Bouchitté, G., and Buttazzo, G. Characterization of optimal shapes and masses through Monge-Kantorovich equation. *J. Eur. Math. Soc. (JEMS) 3*, 2 (2001), 139–168.
67. Bouchitté, G., Buttazzo, G., and Seppecher, P. Shape optimization solutions via Monge-Kantorovich equation. *C. R. Acad. Sci. Paris Sér. I Math. 324*, 10 (1997), 1185–1191.
68. Brascamp, H. J., and Lieb, E. H. On extensions of the Brunn-Minkowski and Prékopa-Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. *J. Functional Analysis 22*, 4 (1976), 366–389.
69. Brenier, Y. Décomposition polaire et réarrangement monotone des champs de vecteurs. *C. R. Acad. Sci. Paris Sér. I Math. 305*, 19 (1987), 805–808.
70. Brenier, Y. The least action principle and the related concept of generalized flows for incompressible perfect fluids. *J. Amer. Math. Soc. 2*, 2 (1989), 225–255.
71. Brenier, Y. The least action principle and the related concept of generalized flows for incompressible perfect fluids. *J. Amer. Math. Soc. 2*, 2 (1989), 225–255.
72. Brenier, Y. Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math. 44*, 4 (1991), 375–417.
73. Brenier, Y. The dual least action problem for an ideal, incompressible fluid. *Arch. Rational Mech. Anal. 122*, 4 (1993), 323–351.
74. Brenier, Y. A homogenized model for vortex sheets. *Arch. Rational Mech. Anal. 138*, 4 (1997), 319–353.
75. Brenier, Y. Minimal geodesics on groups of volume-preserving maps and generalized solutions of the Euler equations. *Comm. Pure Appl. Math. 52*, 4 (1999), 411–452.
76. Brenier, Y. Minimal geodesics on groups of volume-preserving maps and generalized solutions of the Euler equations. *Comm. Pure Appl. Math. 52*, 4 (1999), 411–452.
77. Brenier, Y. Extended Monge-Kantorovich theory. In *Optimal transportation and applications (Martina Franca, 2001)*, vol. 1813 of *Lecture Notes in Math.* Springer, Berlin, 2003, pp. 91–121.
78. Brézis, H. *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*. North-Holland Publishing Co., Amsterdam, 1973. North-Holland Mathematics Studies, No. 5. Notas de Matemática (50).
79. Brézis, H. *Analyse fonctionnelle*. Masson, Paris, 1983. Théorie et applications. [Functional analysis, Theory and applications].
80. Brézis, H., and Lieb, E. Sobolev inequalities with a remainder term. *J. Funct. Anal. 62* (1985), 73–86.

81. Burago, D., Burago, Y., and Ivanov, S. *A course in metric geometry*, vol. 33 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2001. A list of errata is available online at `http://www.pdmi.ras.ru/staff/burago.html`.
82. Burago, Y. D., and Zalgaller, V. A. *Geometric inequalities*. Springer-Verlag, Berlin, 1988. Translated from the Russian by A. B. Sosinskiĭ, Springer Series in Soviet Mathematics.
83. Buttazzo, G., Giaquinta, M., and Hildebrandt, S. *One-dimensional variational problems*, vol. 15 of *Oxford Lecture Series in Mathematics and its Applications*. The Clarendon Press Oxford University Press, New York, 1998. An introduction.
84. Buttazzo, G., Santambrogio, F., and Brancolini, A. Path functionals over Wasserstein spaces. Preprint, 2004.
85. Cabré, X. Nondivergent elliptic equations on manifolds with nonnegative curvature. *Comm. Pure Appl. Math. 50*, 7 (1997), 623–665.
86. Cabré, X. The isoperimetric inequality via the ABP method. preprint, 2005.
87. Caffarelli, L. A. Interior $W^{2,p}$ estimates for solutions of the Monge-Ampère equation. *Ann. of Math. (2) 131*, 1 (1990), 135–150.
88. Caffarelli, L. A. Boundary regularity of maps with convex potentials. *Comm. Pure Appl. Math. 45*, 9 (1992), 1141–1151.
89. Caffarelli, L. A. The regularity of mappings with a convex potential. *J. Amer. Math. Soc. 5*, 1 (1992), 99–104.
90. Caffarelli, L. A. Boundary regularity of maps with convex potentials. II. *Ann. of Math. (2) 144*, 3 (1996), 453–496.
91. Caffarelli, L. A. Monotonicity properties of optimal transportation and the FKG and related inequalities. *Comm. Math. Phys. 214*, 3 (2000), 547–563. Erratum in *Comm. Math. Phys.* 225 (2002), no. 2, 449–450.
92. Caffarelli, L. A., and Cabré, X. *Fully nonlinear elliptic equations*, vol. 43 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI, 1995.
93. Caffarelli, L. A., Feldman, M., and McCann, R. J. Constructing optimal maps for Monge's transport problem as a limit of strictly convex costs. *J. Amer. Math. Soc. 15*, 1 (2002), 1–26 (electronic).
94. Cannarsa, P., and Sinestrari, C. *Semiconcave functions, Hamilton-Jacobi equations, and optimal control*. Progress in Nonlinear Differential Equations and their Applications, 58. Birkhäuser Boston Inc., Boston, MA, 2004.
95. Carlen, E. Behind the wave function: stochastic mechanics today. In *Proceedings of the Joint Concordia Sherbrooke Seminar Series on Functional Integration Methods in Stochastic Quantum Mechanics (Sherbrooke, PQ and Montreal, PQ, 1987)* (1991), no. 25, pp. 141–156.
96. Carlen, E., and Soffer, A. Entropy production by block variable summation and central limit theorems. *Comm. Math. Phys. 140* (1991), 339–371.
97. Carlen, E. A., and Gangbo, W. Constrained steepest descent in the 2-Wasserstein metric. *Ann. of Math. (2) 157*, 3 (2003), 807–846.
98. Carlen, E. A., and Gangbo, W. Constrained steepest descent in the 2-Wasserstein metric. *Ann. of Math. (2) 157*, 3 (2003), 807–846.
99. Carlen, E. A., and Gangbo, W. Solution of a model Boltzmann equation via steepest descent in the 2-Wasserstein metric. *Arch. Ration. Mech. Anal. 172*, 1 (2004), 21–64.
100. Carrillo, J. A., DiFrancesco, M., and Toscani, G. Intermediate asymptotics beyond homogeneity and self-similarity: long-time behavior for $u_t = \delta\phi(u_)$. To appear in *Arch. Rational Mech. Anal.*
101. Carrillo, J. A., DiFrancesco, M., and Toscani, G. Strict contractivity of the 2-Wasserstein distance for the porous medium equation by mass centering. Preprint Univ. Autònoma de Barcelona, 2005.
102. Carrillo, J. A., McCann, R. J., and Villani, C. Contraction in the 2-Wasserstein length space and thermalization of granular media. Preprint HYKE 2004-036 at `http://www.hyke.org/preprint.php`. To appear in *Arch. Rational Mech. Anal.*
103. Carrillo, J. A., McCann, R. J., and Villani, C. Kinetic equilibration rates for granular media and related equations: entropy dissipation and mass transportation estimates. *Rev. Mat. Iberoamericana 19*, 3 (2003), 971–1018.
104. Carrillo, J. A., and Toscani, G. Asymptotic L^1-decay of solutions of the porous medium equation to self-similarity. *Indiana Univ. Math. J. 49*, 1 (2000), 113–142.
105. Carrillo, J. A., and Vázquez, J. L. Fine asymptotics for fast diffusion equations. *Comm. Partial Differential Equations 28*, 5-6 (2003), 1023–1056.
106. Cattiaux, P., and Guillin, A. A criterion for Talagrand's quadratic transportation cost inequality. Preprint, available online via `www.ceremade.dauphine.fr/ guillin/index3.html`, 2004.

107. Cattiaux, P., and Guillin, A. Probabilistic approach for granular media equations in the non uniformly convex case. Preprint, 2006.
108. Chavel, I. *Riemannian geometry—a modern introduction*, vol. 108 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1993.
109. Cheeger, J. Differentiability of Lipschitz functions on metric measure spaces. *Geom. Funct. Anal. 9*, 3 (1999), 428–517.
110. Cheeger, J., and Colding, T. H. On the structure of spaces with Ricci curvature bounded below. I. *J. Differential Geom. 46*, 3 (1997), 406–480.
111. Cheeger, J., and Colding, T. H. On the structure of spaces with Ricci curvature bounded below. III. *J. Differential Geom. 54*, 1 (2000), 37–74.
112. Christensen, J. P. R. Measure theoretic zero sets in infinite dimensional spaces and applications to differentiability of Lipschitz mappings. *Publ. Dép. Math. (Lyon) 10*, 2 (1973), 29–39. Actes du Deuxième Colloque d'Analyse Fonctionnelle de Bordeaux (Univ. Bordeaux, 1973), I, pp. 29–39.
113. Clarke, F. H. *Methods of dynamic and nonsmooth optimization*, vol. 57 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1989.
114. Cordero-Erausquin, D. Inégalité de Prékopa-Leindler sur la sphère. *C. R. Acad. Sci. Paris Sér. I Math. 329*, 9 (1999), 789–792.
115. Cordero-Erausquin, D. Some applications of mass transport to Gaussian-type inequalities. *Arch. Ration. Mech. Anal. 161*, 3 (2002), 257–269.
116. Cordero-Erausquin, D. Non-smooth differential properties of optimal transport. In *Recent advances in the theory and applications of mass transport*, vol. 353 of *Contemp. Math.* Amer. Math. Soc., Providence, RI, 2004, pp. 61–71.
117. Cordero-Erausquin, D., McCann, R., and Schmuckenschlaeger, M. Prékopa-leindler type inequalities on riemannian manifolds, jacobi fields, and optimal transport. To appear in *Ann. Fac. Sci. Toulouse Math. (6)*, available online via `http://www.math.toronto.edu/mccann/`.
118. Cordero-Erausquin, D., McCann, R. J., and Schmuckenschläger, M. A Riemannian interpolation inequality à la Borell, Brascamp and Lieb. *Invent. Math. 146*, 2 (2001), 219–257.
119. Cordero-Erausquin, D., Nazaret, B., and Villani, C. A mass-transportation approach to sharp Sobolev and Gagliardo-Nirenberg inequalities. *Adv. Math. 182*, 2 (2004), 307–332.
120. Csiszar, I. Information-type measures of difference of probability distributions and indirect observations. *Stud. Sci. Math. Hung. 2* (1967), 299–318.
121. Cullen, M. *A mathematical theory of large-scale atmosphere/ocean flow*. World Scientific, 2006.
122. Cullen, M., Douglas, R., Roulstone, I., and Sewell, M. Generalised semi-geostrophic theory on a sphere. Preprint, available online via `http://users.math.uni-potsdam.de/ sreich/NETWORK/publications.html`, 2003.
123. Cullen, M., and Gangbo, W. A variational approach for the 2-dimensional semi-geostrophic shallow water equations. *Arch. Ration. Mech. Anal. 156*, 3 (2001), 241–273.
124. Cullen, M. J. P., and Douglas, R. J. Applications of the Monge-Ampère equation and Monge transport problem to meteorology and oceanography. In *Monge Ampère equation: applications to geometry and optimization (Deerfield Beach, FL, 1997)*. Amer. Math. Soc., Providence, RI, 1999, pp. 33–53.
125. Cullen, M. J. P., and Purser, R. J. Properties of the Lagrangian semigeostrophic equations. *J. Atmospheric Sci. 46*, 17 (1989), 2684–2697.
126. Dacorogna, B., and Moser, J. On a partial differential equation involving the Jacobian determinant. *Ann. Inst. H. Poincaré Anal. Non Linéaire 7*, 1 (1990), 1–26.
127. Davies, E. B. *Heat kernels and spectral theory*, vol. 92 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1989.
128. de Acosta, A. Invariance principles in probability for triangular arrays of B-valued random vectors and some applications. *Ann. Probab. 10*, 2 (1982), 346–373.
129. De Giorgi, E. Sulla differenziabilità e l'analiticità delle estremali degli integrali multipli regolari. *Mem. Accad. Sci. Torino. Cl. Sci. Fis. Mat. Nat. (3) 3* (1957), 25–43.
130. De Pascale, L., and Pratelli, A. Sharp summability for Monge transport density via interpolation. *ESAIM Control Optim. Calc. Var. 10*, 4 (2004), 549–552 (electronic).
131. Del Pino, M., and Dolbeault, J. Best constants for Gagliardo-Nirenberg inequalities and applications to nonlinear diffusions. *J. Math. Pures Appl. (9) 81*, 9 (2002), 847–875.
132. Delanoe, P., and Loeper, G. Gradient estimates for potentials of invertible gradient-mappings on the sphere. Preprint, available online via `math1.unice.fr/ loeper/publications.htm`.
133. Dellacherie, C. Ensembles analytiques. Théorèmes de séparation et applications. In *Séminaire de Probabilités, IX (Seconde Partie, Univ. Strasbourg, Strasbourg, années universitaires 1973/1974 et 1974/1975)*. Springer, Berlin, 1975, pp. 336–372. Lecture Notes in Math., Vol. 465.

134. Delon, J. Personal communication.
135. Demange, J. Porous media equation and Sobolev inequalities under negative curvature. To appear in *Bull. Sci. Mathématiques*.
136. Demange, J. From porous media equations to generalized Sobolev inequalities on a Riemannian manifold. Preprint, available online at `http://www.lsp.ups-tlse/Fp/Demange`, 2004.
137. Demange, J. Improved Sobolev inequalities under positive curvature. Preprint, 2004.
138. Denzler, J., and McCann, R. J. Fast diffusion to self-similarity: complete spectrum, long-time asymptotics, and numerology. *Arch. Ration. Mech. Anal. 175*, 3 (2005), 301–342.
139. Desvillettes, L., and Villani, C. On the spatially homogeneous Landau equation for hard potentials. II. H-theorem and applications. *Comm. Partial Differential Equations 25*, 1-2 (2000), 261–298.
140. Desvillettes, L., and Villani, C. On the trend to global equilibrium in spatially inhomogeneous entropy-dissipating systems: the linear Fokker-Planck equation. *Comm. Pure Appl. Math. 54*, 1 (2001), 1–42.
141. Djellout, H., Guillin, A., and Wu, L. Transportation cost-information inequalities and applications to random dynamical systems and diffusions. *Ann. Probab. 32*, 3B (2004), 2702–2732.
142. do Carmo, M. P. *Riemannian geometry*. Mathematics: Theory & Applications. Birkhäuser Boston Inc., Boston, MA, 1992. Translated from the second Portuguese edition by Francis Flaherty.
143. Dobrushin, L. Prescribing a system of random variables by conditional distributions. *Theor. Prob. Appl. 15* (1970), 458–486.
144. Dobrušin, R. L. Vlasov equations. *Funktsional. Anal. i Prilozhen. 13*, 2 (1979), 48–58, 96.
145. Dudley, R. *Probabilities and metrics : Convergence of laws on metric spaces, with a view to statistical testing*. Aarhus Universitet, 1976.
146. Dudley, R. M. *Uniform central limit theorems*, vol. 63 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1999.
147. Dudley, R. M. *Real analysis and probability*, vol. 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2002. Revised reprint of the 1989 original.
148. Dupin, C. Applications de la géométrie et de la mécanique. Re-edition by Bachelier, Paris (1922).
149. Eckmann, J.-P., and Hairer, M. Uniqueness of the invariant measure for a stochastic PDE driven by degenerate noise. *Comm. Math. Phys. 219*, 3 (2001), 523–565.
150. Einstein, A. Mouvement des particules en suspension dans un fluide au repos, comme conséquence de la théorie cinétique moléculaire de la chaleur. *Annalen der Physik 17* (1905), 549–560. Trad. française.
151. Ekeland, I., and Témam, R. *Convex analysis and variational problems*, english ed., vol. 28 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1999. Translated from the French.
152. Eliassen, A. The quasi-static equations of motion. *Geofys. Publ. 17*, 3 (1948).
153. Evans, L. C. *Partial differential equations*. American Mathematical Society, Providence, RI, 1998.
154. Evans, L. C. Partial differential equations and Monge-Kantorovich mass transfer. In *Current developments in mathematics, 1997 (Cambridge, MA)*. Int. Press, Boston, MA, 1999, pp. 65–126.
155. Evans, L. C., Feldman, M., and Gariepy, R. F. Fast/slow diffusion and collapsing sandpiles. *J. Differential Equations 137*, 1 (1997), 166–209.
156. Evans, L. C., and Gariepy, R. F. *Measure theory and fine properties of functions*. CRC Press, Boca Raton, FL, 1992.
157. Evans, L. C., and Gomes, D. Effective Hamiltonians and averaging for Hamiltonian dynamics. I. *Arch. Ration. Mech. Anal. 157*, 1 (2001), 1–33.
158. Evans, L. C., and Gomes, D. Effective Hamiltonians and averaging for Hamiltonian dynamics. II. *Arch. Ration. Mech. Anal. 161*, 4 (2002), 271–305.
159. Evans, L. C., and Rezakhanlou, F. A stochastic model for growing sandpiles and its continuum limit. *Comm. Math. Phys. 197*, 2 (1998), 325–345.
160. Fathi, A. Weak KAM theorem in Lagrangian dynamics. Lecture notes, preliminary version, 2001.
161. Fathi, A., and Figalli, A. Optimal transportation on manifolds. Preprint, 2006.
162. Feldman, M., and McCann, R. J. Uniqueness and transport density in Monge's mass transportation problem. *Calc. Var. Partial Differential Equations 15*, 1 (2002), 81–113.
163. Feng, J., and Kurtz, T. G. Large deviations for stochastic processes. Research monograph. Available online at `http://www.math.umass.edu/~feng/Research.html`.
164. Fernique, X. Sur le théorème de Kantorovitch-Rubinstein dans les espaces polonais. In *Seminar on Probability, XV (Univ. Strasbourg, Strasbourg, 1979/1980) (French)*, vol. 850 of *Lecture Notes in Math.* Springer, Berlin, 1981, pp. 6–10.
165. Feyel, D., and Üstünel, A. S. Measure transport on Wiener space and the Girsanov theorem. *C. R. Math. Acad. Sci. Paris 334*, 11 (2002), 1025–1028.
166. Feyel, D., and Üstünel, A. S. Monge-Kantorovitch measure transportation and Monge-Ampère equation on Wiener space. *Probab. Theory Related Fields 128*, 3 (2004), 347–385.

167. Feyel, D., and Üstünel, A. S. Monge-Kantorovitch measure transportation, Monge-Ampère equation and the Itô calculus. In *Stochastic analysis and related topics in Kyoto*, vol. 41 of *Adv. Stud. Pure Math.* Math. Soc. Japan, Tokyo, 2004, pp. 49–74.
168. Feyel, D., and Üstünel, A. S. The strong solution of the Monge-Ampère equation on the Wiener space for log-concave densities. *C. R. Math. Acad. Sci. Paris 339*, 1 (2004), 49–53.
169. Forni, G., and Mather, J. N. *Action minimizing orbits in Hamiltonian systems. In Transition to chaos in classical and quantum mechanics*, vol. 1589 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1994. Lectures given at the Third C.I.M.E. Session held in Montecatini Terme, July 6–13, 1991, Edited by S. Graffi.
170. Fragalà, I., Gelli, M., and Pratelli, A. Continuity of an optimal transport in Monge problem. Preprint, 2005.
171. Fukaya, K. Collapsing of Riemannian manifolds and eigenvalues of Laplace operator. *Invent. Math. 87*, 3 (1987), 517–547.
172. Fukaya, K. Hausdorff convergence of Riemannian manifolds and its applications. In *Recent topics in differential and analytic geometry*, vol. 18 of *Adv. Stud. Pure Math.* Academic Press, Boston, MA, 1990, pp. 143–238.
173. Gaffke, N., and Rüschendorf, L. On a class of extremal problems in statistics. *Math. Operationsforsch. Statist. Ser. Optim. 12*, 1 (1981), 123–135.
174. Gallay, T., and Wayne, C. E. Global stability of vortex solutions of the two-dimensional Navier-Stokes equation. *Comm. Math. Phys. 255*, 1 (2005), 97–129.
175. Gallot, S., Hulin, D., and Lafontaine, J. *Riemannian geometry*, second ed. Universitext. Springer-Verlag, Berlin, 1990.
176. Gangbo, W. An elementary proof of the polar factorization of vector-valued functions. *Arch. Rational Mech. Anal. 128*, 4 (1994), 381–399.
177. Gangbo, W. The Monge mass transfer problem and its applications. In *Monge Ampère equation: applications to geometry and optimization (Deerfield Beach, FL, 1997)*, vol. 226 of *Contemp. Math.* Amer. Math. Soc., Providence, RI, 1999, pp. 79–104.
178. Gangbo, W., and McCann, R. J. Optimal maps in Monge's mass transport problem. *C. R. Acad. Sci. Paris Sér. I Math. 321*, 12 (1995), 1653–1658.
179. Gangbo, W., and McCann, R. J. The geometry of optimal transportation. *Acta Math. 177*, 2 (1996), 113–161.
180. Gangbo, W., and Święch, A. Optimal maps for the multidimensional Monge-Kantorovich problem. *Comm. Pure Appl. Math. 51*, 1 (1998), 23–45.
181. Gardner, R. The Brunn-Minkowski inequality. *Bull. Amer. Math. Soc. (N.S.) 39*, 3 (2002), 355–405.
182. Gentil, I. Ultracontractive bounds on Hamilton-Jacobi solutions. *Bull. Sci. Math. 126*, 6 (2002), 507–524.
183. Giacomin, G. On stochastic domination in the Brascamp-Lieb framework. *Math. Proc. Cambridge Philos. Soc. 134*, 3 (2003), 507–514.
184. Gianazzo, U., Savaré, G., and Toscani, G. A fourth order nonlinear pde as gradient flow of the Fisher information in Wasserstein spaces. Preprint, Università di Pavia, 2004.
185. Gilbarg, D., and Trudinger, N. S. *Elliptic partial differential equations of second order*. Classics in Mathematics. Springer-Verlag, Berlin, 2001. Reprint of the 1998 edition.
186. Gini, C. Sulla misura della concentrazione e della variabilità del caratteri. *Atti del R. Instituto Veneto 73* (1914), 1913–1914.
187. Goudon, T., Junca, S., and Toscani, G. Fourier-based distances and Berry-Esseen like inequalities for smooth densities. *Monatsh. Math. 135*, 2 (2002), 115–136.
188. Gozlan, N. *Principe conditionnel de Gibbs pour des contraintes fines approchées et Inégalités de Transport (in french)*. PhD thesis, Univ. Paris X - Nanterre, 2005.
189. Gozlan, N., and Léonard, C. A large deviation approach to some transportation cost inequalities. Preprint, 2005.
190. Greene, R. E., and Shiohama, K. Diffeomorphisms and volume-preserving embeddings of noncompact manifolds. *Trans. Amer. Math. Soc. 255* (1979), 403–414.
191. Grigor′yan, A. Analytic and geometric background of recurrence and non-explosion of the Brownian motion on Riemannian manifolds. *Bull. Amer. Math. Soc. (N.S.) 36*, 2 (1999), 135–249.
192. Grigor′yan, A. A. The heat equation on noncompact Riemannian manifolds. *Mat. Sb. 182*, 1 (1991), 55–87.
193. Gromov, M. Paul Lévy's isoperimetric inequality. Preprint IHES, 1980.
194. Gromov, M. Groups of polynomial growth and expanding maps. *Inst. Hautes Études Sci. Publ. Math.*, 53 (1981), 53–73.
195. Gromov, M. Sign and geometric meaning of curvature. *Rend. Sem. Mat. Fis. Milano 61* (1991), 9–123 (1994).

196. Gromov, M. *Metric structures for Riemannian and non-Riemannian spaces*, vol. 152 of *Progress in Mathematics*. Birkhäuser Boston Inc., Boston, MA, 1999. Based on the 1981 French original [MR0682063 (85e:53051)], With appendices by M. Katz, P. Pansu and S. Semmes, Translated from the French by Sean Michael Bates.
197. Gromov, M., and Milman, V. D. A topological application of the isoperimetric inequality. *Amer. J. Math. 105*, 4 (1983), 843–854.
198. Gross, L. Logarithmic Sobolev inequalities. *Amer. J. Math. 97* (1975), 1061–1083.
199. Gross, L. Logarithmic Sobolev inequalities and contractivity properties of semigroups. In *Dirichlet forms (Varenna, 1992)*. Springer, Berlin, 1993, pp. 54–88. Lecture Notes in Math., 1563.
200. Grove, K., and Petersen, P. Manifolds near the boundary of existence. *J. Differential Geom. 33*, 2 (1991), 379–394.
201. Guittet, K. On the time-continuous mass transport problem and its approximation by augmented Lagrangian techniques. *SIAM J. Numer. Anal. 41*, 1 (2003), 382–399 (electronic).
202. Gutiérrez, C. E. *The Monge-Ampère equation*, vol. 44 of *Progress in Nonlinear Differential Equations and their Applications*. Birkhäuser Boston Inc., Boston, MA, 2001.
203. Hackenbroch, W., and Thalmaier, A. *Stochastische Analysis*. Mathematische Leitfäden. [Mathematical Textbooks]. B. G. Teubner, Stuttgart, 1994. Eine Einführung in die Theorie der stetigen Semimartingale. [An introduction to the theory of continuous semimartingales].
204. Hairer, M., and Mattingly, J. Ergodicity of the 2D Navier-Stokes equations with degenerate stochastic forcing. To appear in *Ann. of Math.*
205. Hairer, M., and Mattingly, J. C. Ergodic properties of highly degenerate 2D stochastic Navier-Stokes equations. *C. R. Math. Acad. Sci. Paris 339*, 12 (2004), 879–882.
206. Hajłasz, P., and Koskela, P. Sobolev meets Poincaré. *C. R. Acad. Sci. Paris Sér. I Math. 320*, 10 (1995), 1211–1215.
207. Hargé, G. A convex/log-concave correlation inequality for Gaussian measure and an application to abstract Wiener spaces. *Probab. Theory Related Fields 130*, 3 (2004), 415–440.
208. Heinonen, J. *Lectures on analysis on metric spaces*. Universitext. Springer-Verlag, New York, 2001.
209. Heinonen, J., and Koskela, P. Quasiconformal maps in metric spaces with controlled geometry. *Acta Math. 181*, 1 (1998), 1–61.
210. Hérau, F., and Nier, F. Isotropic hypoellipticity and trend to equilibrium for the Fokker-Planck equation with a high-degree potential. *Arch. Ration. Mech. Anal. 171*, 2 (2004), 151–218.
211. Hohloch, S. *Optimale Massebewegung im Monge-Kantorovich-Transportproblem, diploma thesis (in German)*. PhD thesis, Freiburg i.Br., 2002.
212. Holley, R., and Stroock, D. Logarithmic Sobolev inequalities and stochastic Ising models. *J. Statist. Phys. 46*, 5–6 (1987), 1159–1194.
213. Hoskins, B. Atmospheric frontogenesis models: some solutions. *Q.J.R. Met. Soc. 97* (1971), 139–153.
214. Hoskins, B. The geostrophic momentum approximation and the semi-geostrophic equations. *J. Atmosph. Sciences 32* (1975), 233–242.
215. Hoskins, B. The mathematical theory of frontogenesis. *Ann. Rev. of Fluid Mech. 14* (1982), 131–151.
216. Hsu, E. P. *Stochastic analysis on manifolds*, vol. 38 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2002.
217. Huang, C., and Jordan, R. Variational formulation for vlasov-poisson-fokker-planck systems. Preprint.
218. Ichihara, K. Curvature, geodesics and the Brownian motion on a Riemannian manifold. I. Recurrence properties, II. Explosion properties. *Nagoya Math. J. 87* (1982), 101–114; 115–125.
219. Jordan, R., Kinderlehrer, D., and Otto, F. The variational formulation of the Fokker-Planck equation. *SIAM J. Math. Anal. 29*, 1 (1998), 1–17.
220. Kantorovich, L. On an effective method of solving certain classes of extremal problems. *Dokl. Akad. Nauk. USSR 28* (1940), 212–215.
221. Kantorovich, L. On the translocation of masses. *Dokl. Akad. Nauk. USSR 37* (1942), 199–201.
222. Kantorovich, L. On a problem of Monge (in Russian). *Uspekhi Mat. Nauk. 3* (1948), 225–226.
223. Kantorovich, L. Autobiography. In *Nobel Lectures, Economics 1969–1980*, A. Lindbeck, Ed., Les Prix Nobel / Nobel Lectures. World Scientific Publishing Co., Singapore, 1992. Available online at `http://nobelprize.org/economics/laureates/1975/kantorovich-autobio.html`.
224. Keith, S. Modulus and the Poincaré inequality on metric measure spaces. *Math. Z. 245*, 2 (2003), 255–292.
225. Keith, S. A differentiable structure for metric measure spaces. *Adv. Math. 183*, 2 (2004), 271–315.
226. Kellerer, H. G. Duality theorems for marginal problems. *Z. Wahrsch. Verw. Gebiete 67*, 4 (1984), 399–432.
227. Kim, Y. J., and McCann, R. J. Sharp decay rates for the fastest conservative diffusions. *C. R. Math. Acad. Sci. Paris 341*, 3 (2005), 157–162.

228. Knothe, H. Contributions to the theory of convex bodies. *Michigan Math. J. 4* (1957), 39–52.
229. Knott, M., and Smith, C. S. On the optimal mapping of distributions. *J. Optim. Theory Appl. 43*, 1 (1984), 39–49.
230. Knott, M., and Smith, C. S. On a generalization of cyclic monotonicity and distances among random vectors. *Linear Algebra Appl. 199* (1994), 363–371.
231. Kullback, S. A lower bound for discrimination information in terms of variation. *IEEE Trans. Inform. Theory 4* (1967), 126–127.
232. Ledoux, M. On an integral criterion for hypercontractivity of diffusion semigroups and extremal functions. *J. Funct. Anal. 105*, 2 (1992), 444–465.
233. Ledoux, M. Inégalités isopérimétriques en analyse et probabilités. *Astérisque*, 216 (1993), Exp. No. 773, 5, 343–375. Séminaire Bourbaki, Vol. 1992/93.
234. Ledoux, M. Isoperimetry and Gaussian analysis. In *Lectures on probability theory and statistics (Saint-Flour, 1994)*. Springer, Berlin, 1996, pp. 165–294.
235. Ledoux, M. Concentration of measure and logarithmic Sobolev inequalities. In *Séminaire de Probabilités, XXXIII*. Springer, Berlin, 1999, pp. 120–216.
236. Ledoux, M. The geometry of Markov diffusion generators. *Ann. Fac. Sci. Toulouse Math. (6) 9*, 2 (2000), 305–366. Probability theory.
237. Ledoux, M. *The concentration of measure phenomenon*. American Mathematical Society, Providence, RI, 2001.
238. Leindler, L. On a certain converse of Hölder's inequality. In *Linear operators and approximation (Proc. Conf., Oberwolfach, 1971)*. Birkhäuser, Basel, 1972, pp. 182–184. Internat. Ser. Numer. Math., Vol. 20.
239. Li, P., and Yau, S.-T. On the parabolic kernel of the Schrödinger operator. *Acta Math. 156*, 3-4 (1986), 153–201.
240. Liese, F., and Vajda, I. *Convex statistical distances*, vol. 95 of *Teubner-Texte zur Mathematik [Teubner Texts in Mathematics]*. BSB B. G. Teubner Verlagsgesellschaft, Leipzig, 1987. With German, French and Russian summaries.
241. Lions, J.-L. *Quelques méthodes de résolution des problèmes aux limites non linéaires*. Dunod, 1969.
242. Lions, P.-L. Personal communication.
243. Lions, P.-L., Papanicolaou, G., and Varadhan, S. Homogeneization of Hamilton-Jacobi equations. Preprint.
244. Lisini, S. Characterization of absolutely continuous curves in Wasserstein spaces. Preprint, 2005.
245. Loeper, G. $c^{1,\alpha}$ regularity of potentials of optimal transportation. Preprint, available online via `math1.unice.fr/ loeper/publications.htm`, 2005.
246. Loeper, G. Continuity of maps solutions of optimal transportation problems. Preprint, 2006.
247. Lott, J., and Villani, C. Ricci curvature for metric-measure spaces via optimal transport. To appear in *Ann. of Math.* Available online via `http://www.umpa.ens-lyon.fr/~cvillani/`, 2004.
248. Lott, J., and Villani, C. Hamilton-Jacobi semigroup on length spaces and applications to functional inequalities. Work in progress, 2005.
249. Lott, J., and Villani, C. Weak curvature bounds and Poincaré inequalities. Preprint, available online via `http://www.umpa.ens-lyon.fr/~cvillani/`, 2005.
250. Ma, X.-N., Trudinger, N., and Wang, X.-J. Regularity of potential functions of the optimal transportation problem. Preprint, 2005.
251. Maggi, F., and Villani, C. Balls have the worst best Sobolev inequalities. Preprint, available online via `www.umpa.ens-lyon.fr/~cvillani`; to appear in *J. Geom. Anal.*
252. Maggi, F., and Villani, C. Balls have the worst Sobolev inequalities. Part II: variants and extensions. Work in progress.
253. Maggi, F., and Villani, C. Balls have the worst best Sobolev inequalities. *J. Geom. Anal. 15*, 1 (2005), 83–121.
254. Mallows, C. L. A note on asymptotic joint normality. *Ann. Math. Statist. 43* (1972), 508–515.
255. Malrieu, F. Logarithmic Sobolev inequalities for some nonlinear PDE's. *Stochastic Process. Appl. 95*, 1 (2001), 109–132.
256. Maroofi, H. *Applications of the Monge-Kantorovich theory*. PhD thesis, Georgia Tech, 2002.
257. Marton, K. A measure concentration inequality for contracting Markov chains. *Geom. Funct. Anal. 6* (1996), 556–571.
258. Massart, P. Saint-Flour lecture notes. Lecture Notes from the 2003 Saint-Flour Probability Summer School. A preliminary version is available online at `http://www.math.u-psud.fr/ massart/`.
259. Mather, J. N. Minimal measures. *Comment. Math. Helv. 64*, 3 (1989), 375–394.
260. Mather, J. N. Action minimizing invariant measures for positive definite Lagrangian systems. *Math. Z. 207*, 2 (1991), 169–207.
261. Mattingly, J. C., Stuart, A. M., and Higham, D. J. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic Process. Appl. 101*, 2 (2002), 185–232.

262. Maurey, B. Some deviation inequalities. *Geom. Funct. Anal. 1*, 2 (1991), 188–197.
263. Maurey, B. Inégalité de Brunn-Minkowski-Lusternik, et autres inégalités géométriques et fonctionnelles. *Astérisque*, 299 (2005), Exp. No. 928, vii, 95–113. Séminaire Bourbaki. Vol. 2003/2004.
264. Maz'ja, V. G. *Sobolev spaces*. Springer Series in Soviet Mathematics. Springer-Verlag, Berlin, 1985. Translated from the Russian by T. O. Shaposhnikova.
265. McCann, R. *A convexity theory for interacting gases and equilibrium crystals*. PhD thesis, Princeton Univ., 1994.
266. McCann, R. J. Existence and uniqueness of monotone measure-preserving maps. *Duke Math. J. 80*, 2 (1995), 309–323.
267. McCann, R. J. A convexity principle for interacting gases. *Adv. Math. 128*, 1 (1997), 153–179.
268. McCann, R. J. Exact solutions to the transportation problem on the line. *R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci. 455*, 1984 (1999), 1341–1380.
269. McCann, R. J. Polar factorization of maps on Riemannian manifolds. *Geom. Funct. Anal. 11*, 3 (2001), 589–608.
270. McNamara, S., and Young, W. Kinetics of a one-dimensional granular medium. *Phys. Fluids A 5*, 1 (1993), 34–45.
271. Meckes, M. Some remarks on transportation cost and related inequalities. To appear in *GAFA Seminar Proceedings*.
272. Menguy, X. Examples of nonpolar limit spaces. *Amer. J. Math. 122*, 5 (2000), 927–937.
273. Mihalas, D., and Mihalas, B. *Foundations of Radiation Hydrodynamics*. Oxford University Press, 1984.
274. Milman, V. D. A new proof of A. Dvoretzky's theorem on cross-sections of convex bodies. *Funkcional. Anal. i Priložen. 5*, 4 (1971), 28–37.
275. Milman, V. D., and Schechtman, G. *Asymptotic theory of finite-dimensional normed spaces*. Springer-Verlag, Berlin, 1986. With an appendix by M. Gromov.
276. Monge, G. Mémoire sur la théorie des déblais et des remblais. In *Histoire de l'Académie Royale des Sciences de Paris* (1781), pp. 666–704.
277. Moser, J. On Harnack's theorem for elliptic differential equations. *Comm. Pure Appl. Math. 14* (1961), 577–591.
278. Moser, J. A Harnack inequality for parabolic differential equations. *Comm. Pure Appl. Math. 17* (1964), 101–134. Erratum in *Vol. 20 (1967), pp. 231–236*.
279. Moser, J. On the volume elements on a manifold. *Trans. Amer. Math. Soc. 120* (1965), 286–294.
280. Murata, H., and Tanaka, H. An inequality for certain functional of multidimensional probability distributions. *Hiroshima Math. J. 4* (1974), 75–81.
281. Nash, J. Continuity of solutions of parabolic and elliptic equations. *Amer. J. Math. 80* (1958), 931–954.
282. Nelson, E. *Dynamical theories of Brownian motion*. Princeton University Press, Princeton, N.J., 1967. Available online at `http://www.math.princeton.edu/~nelson/books.html`.
283. Nelson, E. The free Markoff field. *J. Functional Analysis 12* (1973), 211–227.
284. Neunzert, H. An introduction to the nonlinear Boltzmann-Vlasov equation. In *Kinetic theories and the Boltzmann equation*, C. Cercignani, Ed., vol. 1048 of *Lecture Notes in Mathematics*. Springer, Berlin, Heidelberg, 1984, pp. 60–110.
285. Øksendal, B. *Stochastic differential equations*, sixth ed. Universitext. Springer-Verlag, Berlin, 2003. An introduction with applications.
286. Osserman, R. The isoperimetric inequality. *Bull. Amer. Math. Soc. 84*, 6 (1978), 1182–1238.
287. Otto, F. Double degenerate diffusion equations as steepest descent. Preprint Univ. Bonn, 1996.
288. Otto, F. Dynamics of labyrinthine pattern formation in magnetic fluids: a mean-field theory. *Arch. Rational Mech. Anal. 141*, 1 (1998), 63–103.
289. Otto, F. Lubrication approximation with prescribed nonzero contact angle. *Comm. Partial Differential Equations 23*, 11-12 (1998), 2077–2164.
290. Otto, F. The geometry of dissipative evolution equations: the porous medium equation. *Comm. Partial Differential Equations 26*, 1-2 (2001), 101–174.
291. Otto, F., and Tzavaras, A. Continuity of velocity gradients in suspensions of rod-like molecules. Preprint, available online via `www.math.wisc.edu/ tzavaras/listsu.html`.
292. Otto, F., and Villani, C. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *J. Funct. Anal. 173*, 2 (2000), 361–400.
293. Otto, F., and Villani, C. Comment on: "Hypercontractivity of Hamilton-Jacobi equations", *j. math. pures appl. (9)* **80** (2001), no. 7, 669–696; by S. G. Bobkov, I. Gentil and M. Ledoux. *J. Math. Pures Appl. (9) 80*, 7 (2001), 697–700.
294. Otto, F., and Westdickenberg, M. Eulerian calculus for the contraction in the Wasserstein distance. Preprint, available online via `www-mathphys.iam.uni-bonn.de/ michael/web/main-de.htm`, 2005.

295. Oxtoby, J. C. Ergodic sets. *Bull. Amer. Math. Soc. 58* (1952), 116–136.
296. Petersen, P. *Riemannian geometry*, vol. 171 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1998.
297. Pinsker, M. *Information and Information Stability of Random Variables and Processes*. Holden-Day, San Francisco, 1964.
298. Pogorelov, A. V. *Monge-Ampère equations of elliptic type*. Groningen, Noordhoff, 1964.
299. Pogorelov, A. V. The Dirichlet problem for the multidimensional analogue of the Monge-Ampère equation. *Dokl. Akad. Nauk SSSR 201* (1971), 790–793. English translation in *Soviet Math. Dokl. 12* (1971), 1727–1731.
300. Pratelli, A. Equivalence between some definitions for the optimal mass transport problem and for the transport density on manifolds. To appear in *Ann. Mat. Pura Appl.* Available online via `www.cvgmt.sns.it`.
301. Pratelli, A. On the equality between Monge's infimum and Kantorovich's minimum in optimal mass transportation. Available online at `http://cvgmt.sns.it/people/pratelli/`, 2004.
302. Pratelli, A. About the sufficiency of the c-cyclical monotonicity for the optimality of transport plans. Preprint, 2006.
303. Prékopa, A. On logarithmic concave measures and functions. *Acta Sci. Math. (Szeged) 34* (1973), 335–343.
304. Prigozhin, L. Variational model of sandpile growth. *European J. Appl. Math. 7*, 3 (1996), 225–235.
305. Pulvirenti, A., and Toscani, G. The theory of the nonlinear Boltzmann equation for Maxwell molecules in Fourier representation. *Ann. Mat. pura ed appl. 171*, 4 (1996), 181–204.
306. Rachev, S., and Rüschendorf, L. *Mass Transportation Problems. Vol. I: Theory, Vol. II: Applications*. Probability and its applications. Springer-Verlag, New York, 1998.
307. Rachev, S. T. *Probability metrics and the stability of stochastic models*. John Wiley & Sons Ltd., Chichester, 1991.
308. Rademacher, H. Über partielle und totale differenzierbarkeit von Funktionen mehrerer Variabeln und über die Transformation der Doppelintegrale. *Math. Ann. 79*, 4 (1919), 340–359.
309. Ramachandran, D., and Rüschendorf, L. A general duality theorem for marginal problems. *Probab. Theory Related Fields 101*, 3 (1995), 311–319.
310. Ramachandran, D., and Rüschendorf, L. Duality and perfect probability spaces. *Proc. Amer. Math. Soc. 124*, 7 (1996), 2223–2228.
311. Rey-Bellet, L., and Thomas, L. E. Exponential convergence to non-equilibrium stationary states in classical statistical mechanics. *Comm. Math. Phys. 225*, 2 (2002), 305–329.
312. Rockafellar, R. T. *Convex analysis*. Princeton University Press, Princeton, NJ, 1997. Reprint of the 1970 original, Princeton Paperbacks.
313. Rosenau, P. Tempered diffusion: A transport process with propagating front and initial delay. *Phys. Rev. A 46* (1992), 7371–7374.
314. Rosenblatt, M. Remarks on a multivariate transformation. *Ann. Math. Statistics 23* (1952), 470–472.
315. Rudin, W. *Real and complex analysis*, third ed. McGraw-Hill Book Co., New York, 1987.
316. Rüschendorf, L. Wasserstein-metric. Historical note. Personal communication.
317. Rüschendorf, L. Sharpness of Fréchet-bounds. *Z. Wahrsch. Verw. Gebiete 57*, 2 (1981), 293–302.
318. Rüschendorf, L. Fréchet-bounds and their applications. In *Advances in probability distributions with given marginals (Rome, 1990)*, vol. 67 of *Math. Appl.* Kluwer Acad. Publ., Dordrecht, 1991, pp. 151–187.
319. Rüschendorf, L. Optimal solutions of multivariate coupling problems. *Appl. Math. (Warsaw) 23*, 3 (1995), 325–338.
320. Rüschendorf, L. On c-optimal random variables. *Statist. Probab. Lett. 27*, 3 (1996), 267–270.
321. Rüschendorf, L., and Rachev, S. T. A characterization of random variables with minimum L^2-distance. *J. Multivariate Anal. 32*, 1 (1990), 48–54. Corrigendum in *J. Multivariate Anal. 34, 1 (1990), p. 156*.
322. Rüschendorf, L., and Rachev, S. T. A characterization of random variables with minimum L^2-distance. *J. Multivariate Anal. 32*, 1 (1990), 48–54. Corrigendum in *Vol. 34, No. 1, p.156*.
323. Rüschendorf, L., and Uckelmann, L. On optimal multivariate couplings. In *Distributions with given marginals and moment problems (Prague, 1996)*. Kluwer Acad. Publ., Dordrecht, 1997, pp. 261–273.
324. Rüschendorf, L., and Uckelmann, L. Numerical and analytical results for the transportation problem of Monge-Kantorovich. *Metrika 51*, 3 (2000), 245–258 (electronic).
325. Rüschendorf, L., and Uckelmann, L. On the n-coupling problem. *J. Multivariate Anal. 81*, 2 (2002), 242–258.
326. Saloff-Coste, L. A note on Poincaré, Sobolev, and Harnack inequalities. *Internat. Math. Res. Notices*, 2 (1992), 27–38.

327. Saloff-Coste, L. *Aspects of Sobolev-type inequalities*, vol. 289 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, Cambridge, 2002.
328. Schachermayer, W., and Teichmann, J. Solution of a problem in Villani's book. To appear in *Proc. AMS*, 2005.
329. Schneider, R. *Convex bodies: the Brunn-Minkowski theory*. Cambridge University Press, Cambridge, 1993.
330. Siburg, K. F. *The principle of least action in geometry and dynamics*, vol. 1844 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2004.
331. Smith, C., and Knott, M. On Hoeffding-Fréchet bounds and cyclic monotone relations. *J. Multivariate Anal. 40*, 2 (1992), 328–334.
332. Spohn, H. *Large scale dynamics of interacting particles*. Texts and Monographs in Physics. Springer-Verlag, Berlin, 1991.
333. Stam, A. Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Inform. Control 2* (1959), 101–112.
334. Stroock, D. W. *An introduction to the analysis of paths on a Riemannian manifold*, vol. 74 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2000.
335. Sturm, K.-T. On the geometry of metric measure spaces. Preprint, available online via `www-wt.iam.uni-bonn.de/ sturm/de/index.html`, 2004.
336. Sturm, K.-T. On the geometry of metric measure spaces. i. To appear in *Acta Mathematica*, after merging with Part II. Available online at `http://www-wt.iam.uni-bonn.de/~sturm/de/`, 2004.
337. Sturm, K.-T. Convex functionals of probability measures and nonlinear diffusions on manifolds. *J. Math. Pures Appl. (9) 84*, 2 (2005), 149–168.
338. Sturm, K.-T. Convex functionals of probability measures and nonlinear diffusions on manifolds. *J. Math. Pures Appl. (9) 84*, 2 (2005), 149–168.
339. Sturm, K.-T. On the geometry of metric measure spaces II. Preprint, 2005.
340. Sturm, K.-T. On the geometry of metric measure spaces. ii. To appear in *Acta Mathematica*, after merging with Part I. Available online at `http://www-wt.iam.uni-bonn.de/~sturm/de/`, 2005.
341. Sturm, K.-T., and von Renesse, M.-K. Transport inequalities, gradient estimates, entropy and Ricci curvature. To appear in *Comm. Pure Appl. Math.*
342. Sudakov, V. Geometric problems in the theory of infinite-dimensional probability distributions. *Proc. Steklov Inst. Math. 141* (1979), 1–178.
343. Sznitman, A. Equations de type de Boltzmann, spatialement homogènes. *Z. Wahrsch. Verw. Gebiete 66* (1984), 559–562.
344. Sznitman, A.-S. Topics in propagation of chaos. In *École d'Été de Probabilités de Saint-Flour XIX—1989*. Springer, Berlin, 1991, pp. 165–251.
345. Szulga, A. On minimal metrics in the space of random variables. *Teor. Veroyatnost. i Primenen. 27*, 2 (1982), 401–405.
346. Talagrand, M. Concentration of measure and isoperimetric inequalities in product spaces. *Inst. Hautes Études Sci. Publ. Math.*, 81 (1995), 73–205.
347. Talagrand, M. New concentration inequalities in product spaces. *Invent. Math. 126*, 3 (1996), 505–563.
348. Talagrand, M. Transportation cost for Gaussian and other product measures. *Geom. Funct. Anal. 6*, 3 (1996), 587–600.
349. Tanaka, H. An inequality for a functional of probability distributions and its application to Kac's one-dimensional model of a Maxwellian gas. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 27* (1973), 47–52.
350. Tanaka, H. Probabilistic treatment of the Boltzmann equation of Maxwellian molecules. *Z. Wahrsch. Verw. Gebiete 46*, 1 (1978/79), 67–105.
351. Thirring, W. *Classical mathematical physics*, third ed. Springer-Verlag, New York, 1997. Dynamical systems and field theories, Translated from the German by Evans M. Harrell, II.
352. Thorisson, H. *Coupling, stationarity, and regeneration*. Probability and its Applications (New York). Springer-Verlag, New York, 2000.
353. Toscani, G. New a priori estimates for the spatially homogeneous Boltzmann equation. *Cont. Mech. Thermodyn. 4* (1992), 81–93.
354. Toscani, G. Sur l'inégalité logarithmique de Sobolev. *C.R. Acad. Sci. Paris, Série I, 324* (1997), 689–694.
355. Toscani, G. Entropy production and the rate of convergence to equilibrium for the Fokker-Planck equation. *Quart. Appl. Math. 57*, 3 (1999), 521–541.
356. Trudinger, N. S. Isoperimetric inequalities for quermassintegrals. *Ann. Inst. H. Poincaré Anal. Non Linéaire 11*, 4 (1994), 411–425.
357. Trudinger, N. S., and Wang, X.-J. On the Monge mass transfer problem. *Calc. Var. PDE 13* (2001), 19–31.

358. Unterreiter, A., Arnold, A., Markowich, P., and Toscani, G. On generalized Csiszár-Kullback inequalities. *Monatsh. Math. 131*, 3 (2000), 235–253.
359. Urbas, J. Mass transfer problems. Lecture notes from a course given in Univ. Bonn, 1997–1998.
360. Urbas, J. On the second boundary value problem for equations of Monge-Ampère type. *J. Reine Angew. Math. 487* (1997), 115–124.
361. Vasershtein, L. N. Markov processes over denumerable products of spaces describing large system of automata. *Problemy Peredači Informacii 5*, 3 (1969), 64–72.
362. Vázquez, J.-L. Smoothing and decay estimates for nonlinear parabolic equations of porous medium type. Lecture Notes, preliminary version of July 2005.
363. Vázquez, J. L. An introduction to the mathematical theory of the porous medium equation. In *Shape optimization and free boundaries (Montreal, PQ, 1990)*. Kluwer Acad. Publ., Dordrecht, 1992, pp. 347–389.
364. Villani, C. A review of mathematical topics in collisional kinetic theory. In *Handbook of mathematical fluid dynamics, Vol. I*. North-Holland, Amsterdam, 2002, pp. 71–305.
365. Villani, C. *Topics in optimal transportation*, vol. 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.
366. Villani, C. Hypocoercivity. Work in progress, 2005.
367. Villani, C. Mathematics of granular materials. Survey paper, available online at `http://www.umpa.ens-lyon.fr/~cvillani`, 2005.
368. von Renesse, M.-K. On local Poincaré via transportation. Preprint, archived at `math.MG/0505588`, 2005.
369. Werner, R. The uncertainty relation for joint measurement of position and momentum. Preprint, archived as `quant-ph/0405184`, 2005.

List of Short Statements

List of Figures

C. Villani

UMPA, ENS Lyon
46 allée d'Italie
69364 Lyon Cedex 07
FRANCE

Email: cvillani@umpa.ens-lyon.fr

Webpage: http://www.umpa.ens-lyon.fr/~cvillani