WEIGHTED CSISZÁR-KULLBACK-PINSKER INEQUALITIES AND APPLICATIONS TO TRANSPORTATION INEQUALITIES

FRANÇOIS BOLLEY AND CÉDRIC VILLANI

Abstract. We strengthen the usual Csiszár-Kullback-Pinsker inequality by allowing weights in the total variation norm; admissible weights depend on the decay of the reference probability measure. We use this result to derive transportation inequalities involving Wasserstein distances for various exponents: in particular, we recover the equivalence between a T_1 inequality and the existence of a square-exponential moment. Then we give a variant of the results obtained by Djellout, Guillin and Wu [5] about transportation inequalities for random dynamical systems, in which a sufficient condition is expressed in terms of exponential moments. A result by Blower [1] about the perturbation of a T_2 inequality is also recovered and generalized.

Contents

1. Introduction

Let X be an abstract Polish space, and let $P(X)$ be the set of all Borel probability measures on X; let d be a lower semi-continuous metric on X, and let p belong to $[1, +\infty)$. Whenever μ, ν belong to $P(X)$, we define

- the **Wasserstein distance of order** p between μ and ν by

$$W_p(\mu,\nu) = \inf \left(\iint d(x,y)^p d\pi(x,y) \right)^{1/p}$$

where π runs over the set of probability measures on $X \times X$ with marginals μ and ν;

- the **Kullback information** of μ with respect to ν by

$$H(\mu|\nu) = \int f \,\log f \, d\nu, \qquad f = \frac{d\mu}{d\nu};$$

ENS Lyon, UMPA, 46 allée d'Italie, F-69364 Lyon Cedex 07, FRANCE.
Email addresses: `fbolley@umpa.ens-lyon.fr`, `cvillani@umpa.ens-lyon.fr`.

by convention $H(\mu|\nu) = +\infty$ if μ is not absolutely continuous with respect to ν.

Both objects play an important role in a number of problems in probability theory, where they may be encountered under the names of Monge-Kantorovich distances, or minimal distances, and relative entropy, or relative H functional. More information can be found, together with many references, in [11]. For various purposes it is of interest to investigate whether they can be compared to each other. The most famous such inequality is the **Csiszár-Kullback-Pinsker inequality**, which we shall denote CKP inequality for short: if d is the trivial distance, i.e. $d(x,y) = 1_{x\neq y}$, then

$$2W_1(\mu,\nu) = \|\mu - \nu\|_{TV} \leq \sqrt{2H(\mu|\nu)},$$

where "TV" stands for the total variation norm.

Another class of inequalities which has been studied at length is encountered under the names of **Talagrand inequalities**, **transportation inequalities**, or **transportation cost-information inequalities**; we shall just denote it by T_p. By definition, a reference probability measure ν satisfies the $T_p(\lambda)$ inequality for some $\lambda > 0$ if

$$\forall \mu \in P(X), \quad W_p(\mu,\nu) \leq \sqrt{\frac{2H(\mu|\nu)}{\lambda}};$$

and it satisfies T_p if it satisfies $T_p(\lambda)$ for some $\lambda > 0$. In particular, CKP inequality means that *any* reference probability measure satisfies $T_1(4)$ when d is the trivial distance.

We note right away that $W_p \leq W_{p'}$ for $p \leq p'$, so that T_p inequalities become stronger and stronger as p becomes larger. The cases $p = 1$ and $p = 2$ are of particular interest.

The study of T_p inequalities is a rather old topic [9], which recently received a new impulse. First, it was pointed out by Marton [7] and Talagrand [10] that these inequalities are a handy tool in the study of *concentration of measure* [6]; in particular, Talagrand showed how to take advantage of the good tensorization properties of inequality T_2, to establish concentration in product spaces. At the same time, he established the validity of T_2 for the Gaussian measure, which justifies the terminology of "Talagrand inequalities". On the other hand, recent developments of the theory of optimal transportation led to new connections between these inequalities and other classes of functional inequalities with a geometric content, in particular **logarithmic Sobolev inequalities**. For instance, the main result in [8] is that *a logarithmic Sobolev inequality implies a T_2 inequality* (and the converse is also true under some convexity assumption). Various proofs and variants of these results, together with a detailed discussion, can be found in [2, 8, 11].

On the other hand, the works by Bobkov and Götze [3], and Djellout, Guillin and Wu [5] suggest that there is still room for investigation in an abstract Polish space setting, without any underlying geometric structure. More precisely, given a reference probability measure ν, one of the main results proven in these references is the equivalence between

(1) ν satisfies a T_1 inequality;

(2) there exists λ such that $\int e^{t(f(x)-\int f(x)\,d\nu(x))}\,d\nu(x) \leq e^{\frac{t^2}{2\lambda}}$ for any real t and Lipschitz function f with Lipschitz seminorm 1;

(3) ν admits a square-exponential moment, i.e. $\int e^{\alpha d(x,y)^2}\, d\nu(x)$ is finite for some $\alpha > 0$ and some (and thus any) y.

Notice how tractable is this criterium for T_1: for instance, the validity of a logarithmic Sobolev inequality depends on subtle properties of the reference measure, which imply not only the existence of a square-exponential moment, but also – among other features which are still poorly identified – strict positivity, in a quantitative way which has not been made precise so far (see however [4] for important progress in that direction). Djellout, Guillin and Wu explored various applications of their result, including T_1 inequalities in path space for solutions of stochastic differential equations, or T_1 inequalities in large dimension for random dynamical systems under adequate assumptions of weak dependence.

The purpose of this paper is twofold.

On one hand, we shall establish a generalization of the CKP inequality, allowing for a weight in the total variation. How much weight is allowed will depend on the decay of the reference measure. In that generalization, the optimal constant 4 will be lost, but this will be more than compensated by the gain of precision brought by the weight. In view of the large range of applications of the usual CKP inequality, we do hope that this generalization can be of interest in various contexts.

On the other hand, we shall point out that, instead of considering CKP inequality as just a particular case of T_1, it is possible to establish many general comparison results between W_p and H by studying the weighted CKP inequality. In particular, we shall recover in a straightforward way (and with improved constants) the above-mentionned result according to which a square-exponential moment implies T_1. Then we shall establish a variant of the result by Djellout, Wu and Guillin [5] about random dynamical systems, in which assumptions are only expressed in terms of exponential moments. Not only are these conditions easier to check, but they also allow for more generality. Also, we shall establish weakened versions of T_1 and T_2 inequalities, in which the square-root on the right-hand side is replaced by a combination of powers, and which are satisfied with quite a bit of generality, under just decay assumptions on the reference measure. Among them is a generalization of a partial result by Blower [1] about the perturbation of T_2 inequalities.

The plan of the paper is as follows. In section 2, we state our weighted CKP inequality and derive from it various applications to the study of T_p inequalities and their variants. In section 3, we give a detailed proof of the weighted CKP inequality and in section 4 we show how our results can be applied to the study of discrete-time processes. Finally, in the appendix, we give another proof of the equivalence between a T_1 inequality and the existence of a square-exponential moment.

2. Main results

Working in a Polish space is a natural assumption when handling Wasserstein distances, because it is sufficient to derive all the well-known and useful properties of these distances, in particular their relation with the weak topology [11]. However, for all the results in this section, no use will be made of completeness or separability, and so we state the results with more generality.

In the sequel, the notation $\varphi(\mu-\nu)$ is a shorthand for the signed measure $\varphi\mu - \varphi\nu$.

Theorem 1 (weighted CKP inequalities)**.** *Let X be a measurable space, let μ, ν be two probability measures on X, and let φ be a nonnegative measurable function on X. Then*

$$\text{(i)} \qquad \|\varphi\,(\mu-\nu)\|_{TV} \;\le\; \left(\frac{3}{2} + \log \int e^{2\varphi(x)}\,d\nu(x)\right)\left(\sqrt{H(\mu|\nu)} + \frac{1}{2}H(\mu|\nu)\right);$$

$$\text{(ii)} \qquad \|\varphi\,(\mu-\nu)\|_{TV} \;\le\; \sqrt{2}\left(1 + \log \int e^{\varphi(x)^2}\,d\nu(x)\right)^{1/2}\sqrt{H(\mu|\nu)}.$$

Remarks 2. 1. The assumption $\displaystyle\int_X e^{\varphi^2}\,d\nu < +\infty$ is always stronger than the assumption $\int_X e^{2\varphi}\,d\nu < +\infty$, so the inequality (i) above always applies in more generality than (ii). Further note that if we choose $\varphi \equiv 1$ in (ii), we recover the usual CKP inequality

$$\|\mu-\nu\|_{TV} \;\le\; c\sqrt{H(\mu|\nu)}$$

with the non-optimal constant $c=2$ instead of $\sqrt{2}$. This shows that the constants on the right-hand side of (ii) cannot be improved by more than a factor $\sqrt{2}$. Although we worked quite a bit to decrease this numerical constant, it is likely that one can still do better, at least by replacing $\int e^{\varphi^2}$ with $\int e^{\lambda\varphi^2}$. Note though that the optimal constant $\sqrt{2}$ can be recovered by writing our proof again in the particular case $\varphi \equiv 1$, as it shall be pointed out in section 3.

2. Let us discuss very briefly the sharpness of the orders of magnitude in the above inequalities. When μ is very close to ν, the Kullback information can be approximated by a weighted squared L^2 norm, which shows that it is natural to expect a term in $\sqrt{H(\mu|\nu)}$ (as opposed to another power of H) in the right-hand side. On the other hand, consider the situation when $X=\mathbb{R}^n$, and the reference measure ν is the standard Gaussian distribution; choose $\varphi(x) = \delta|x|$ for $\delta < 1/\sqrt{2}$. Then the left-hand side of inequality (ii) will be typically $O(\sqrt{n})$ as $n\to\infty$, while the right-hand side will be typically $O(n)$. If $\varphi(x) = \delta\sum|x_i|/\sqrt{n}$, then the left-hand side will be typically $O(n)$, while the right-hand side will be typically $O(n^{3/2})$. These examples suggest that Theorem 1 still leaves room for improvement for problems set in large dimension. As we shall see in Section 4, this loss of a $O(\sqrt{n})$ factor will put limitation on the validity of measure concentration inequalities that can be deduced from Theorem 1 in large dimension.

We postpone the proof of Theorem 1 to the next section, and now list two consequences.

Corollary 3. *Let X be a measurable space equipped with a measurable distance d, let $p\ge 1$ and let ν be a probability measure on X. Assume that there exist $x_0\in X$ and $\alpha>0$ such that $\displaystyle\int e^{\alpha d(x_0,x)^p}\,d\nu(x)$ is finite. Then*

$$\forall \mu \in P(X), \qquad W_p(\mu,\nu) \;\le C\Big[H(\mu|\nu)^{\frac{1}{p}} + \Big(\frac{H(\mu|\nu)}{2}\Big)^{\frac{1}{2p}}\Big],$$

where

$$C := 2 \inf_{x_0 \in X,\, \alpha>0} \left(\frac{1}{\alpha} \Big(\frac{3}{2} + \log \int e^{\alpha d(x_0,x)^p} \, d\nu(x) \Big) \right)^{\frac{1}{p}} < +\infty.$$

Corollary 4. *Let X be a measurable space equipped with a measurable distance d, let $p \geq 1$ and let ν be a probability measure on X. Assume that there exist $x_0 \in X$ and $\alpha > 0$ such that $\int e^{\alpha d(x_0,x)^{2p}} \, d\nu(x)$ is finite. Then*

$$\forall \mu \in P(X), \qquad W_p(\mu,\nu) \leq C\, H(\mu|\nu)^{\frac{1}{2p}},$$

where

$$C := 2 \inf_{x_0 \in X,\, \alpha>0} \left(\frac{1}{2\alpha} \Big(1 + \log \int e^{\alpha d(x_0,x)^{2p}} \, d\nu(x) \Big) \right)^{\frac{1}{2p}} < +\infty.$$

Particular case 5. When X is bounded, a simpler bound holds:

$$\forall \mu \in P(X), \qquad W_p(\mu,\nu) \leq 2^{\frac{1}{2p}} \mathrm{diam}(X)\, H(\mu|\nu)^{\frac{1}{2p}},$$

where $\mathrm{diam}(X) := \sup\{d(x,y);\ x,y \in X\}$.

Since the proofs of these results are very similar, we only give the proof of Corollary 4.

Proof of Corollary 4. On one hand it is known [11, Proposition 7.10] that

$$W_p^p(\mu,\nu) \leq 2^{p-1} \| d(x_0,.)^p\, (\mu-\nu)\|_{TV};$$

on the other hand the second part of Theorem 1 yields

$$\Big\| \sqrt{\alpha}\, d(x_0,.)^p\, (\mu-\nu) \Big\|_{TV} \leq \sqrt{2} \left(1 + \log \int e^{\alpha d(x_0,x)^{2p}} \, d\nu(x) \right)^{1/2} \sqrt{H(\mu|\nu)}.$$

This concludes the argument. □

We now focus on some particular cases of interest, namely for $p = 1$ and $p = 2$ under assumptions of exponential moments of order 1, 2 and 4.

Corollary 6. *Let X be a measurable space equipped with a measurable distance d, let ν be a reference probability measure on X, and let x_0 be any element of X. Then*

(i) If $\int_X e^{\alpha d(x_0,x)} \, d\nu(x) < +\infty$ for some $\alpha > 0$, then there is a constant C such that

$$\forall \mu \in P(X), \qquad W_1(\mu,\nu) \leq C \left(H(\mu|\nu) + \sqrt{H(\mu|\nu)} \right).$$

(ii) If $\int_X e^{\alpha d(x_0,x)^2} \, d\nu(x) < +\infty$ for some $\alpha > 0$, then there is a constant C such that

$$\forall \mu \in P(X), \qquad W_1(\mu,\nu) \leq C \sqrt{H(\mu|\nu)};$$

$$\forall \mu \in P(X), \qquad W_2(\mu,\nu) \leq C \left[\sqrt{H(\mu|\nu)} + H(\mu|\nu)^{\frac{1}{4}} \right].$$

In particular, ν satisfies T_1.

(iii) If $\int_X e^{\alpha d(x_0,x)^4}\,d\nu(x) < +\infty$ for some $\alpha > 0$, then there is a constant C such that

$$\forall \mu \in P(X), \qquad W_2(\mu,\nu) \le C\,H(\mu|\nu)^{\frac{1}{4}}.$$

Remark 7. Part (ii) of this corollary contains the result that the existence of an exponential moment of order 2 implies a T_1 inequality; according to [3], the converse is true, so this criterion is optimal. To compare these various results in practical situations, it is good to keep in mind the following elementary lemma:

Lemma 8. *Let X be a measurable space equipped with a measurable distance d, let $p \ge 1$ and let ν be a probability measure on X. Then the following three statements are equivalent:*

1. *there exist $x_0 \in X$ and $\alpha > 0$ such that $\int e^{\alpha d(x_0,x)^p}\,d\nu(x)$ is finite;*
2. *for any $x_0 \in X$, there exists $\alpha > 0$ such that $\int e^{\alpha d(x_0,x)^p}\,d\nu(x)$ is finite;*
3. *there exists $\alpha > 0$ such that $\iint e^{\alpha d(x,y)^p}\,d\nu(x)\,d\nu(y)$ is finite.*

Moreover,

$$\inf_{x_0\in X}\int e^{\alpha d(x_0,x)^p}\,d\nu(x) \le \iint e^{\alpha d(x,y)^p}\,d\nu(x)\,d\nu(y) \le \left(\inf_{x_0\in X}\int e^{\alpha 2^{p-1} d(x_0,x)^p}\,d\nu(x)\right)^2.$$

Remark 9. The following two results can be deduced from the equivalence between the existence of an exponential moment of order 2 and a T_1 inequality:

1. Let μ be a probability measure on a Polish space X, satisfying T_1. Then so does any probability measure $\nu = h\mu$, where h is a μ-almost surely bounded measurable function on X.
2. Let μ be a probability measure on $\mathbb{R}^d$ satisfying T_1. Then so does its marginal (via orthogonal projection) on any hyperplane of $\mathbb{R}^d$.

Remark 10. Part (ii) also generalizes the perturbation result proven by Blower, who showed in [1] that an inequality of the form $W_2 \le C(H^{1/2} + H^{1/4})$ holds true when ν is bounded from above and below by constant multiples of a reference measure ν_0 satisfying T_2. In fact, if ν_0 satisfies T_2, then it also satisfies T_1, so it has a finite square-exponential moment, and so does ν if it is bounded above by a constant multiple of ν_0.

Remark 11. Let ν be a reference probability measure having finite exponential moments of order p; how far is it from satisfying T_p? The preceding results indicate that the answer is very different for $p = 1$ and $p = 2$. If T_1 is not satisfied, this means that the decay of ν at infinity is not fast enough, and the T_1 inequality usually fails for *large* values of the Kullback information. On the contrary, if T_2 is not satisfied, this is not necessarily just for a question of fast decay (remember that T_2 implies strict positivity), and the T_2 inequality usually fails for *small* values of the Kullback information. In particular, it is no wonder that we did not manage to recover T_2 inequalities with our arguments taking into account only the decay of ν.

3. PROOF OF THE MAIN INEQUALITIES

We shall now present detailed proofs of the main inequalities in Theorem 1.

Proof of Theorem 1.

Without loss of generality, we assume that μ is absolutely continuous with respect to ν, with density f. We set $u := f - 1$, so that

$$\mu \;=\; (1+u)\nu;$$

we note that $u \geq -1$ and $\displaystyle\int_X u\, d\nu \;=\; 0$. We also define

$$h(v) \;:=\; (1+v)\log(1+v) - v, \qquad v \in [-1, +\infty)$$

so that

$$H(\mu|\nu) \;=\; \int_X h(u)\, d\nu. \tag{3.1}$$

We note that $h \geq 0$.

We start with the **proof of inequality (i)**, splitting the weighted total variation as

$$\int \varphi\, d|\mu - \nu| \;=\; \int \varphi\, |u|\, d\nu \;=\; \int_{\{-1 \leq u \leq 4\}} \varphi\, |u|\, d\nu \;+\; \int_{\{u > 4\}} \varphi\, u\, d\nu. \tag{3.2}$$

We shall estimate both terms separately, first bounding the **first term** ($u \leq 4$) in (3.2). By Cauchy-Schwarz inequality,

$$\int_{u\leq 4} \varphi\, |u|\, d\nu \leq \left(\int_{u\leq 4} \varphi^2\, d\nu\right)^{1/2} \left(\int_{u\leq 4} u^2\, d\nu\right)^{1/2}.$$

On the other hand, from the elementary inequality

$$-1 \leq v \leq 4 \quad \Longrightarrow \quad v^2 \leq 4h(v)$$

(a consequence of the fact that $h(v)/v$ is nondecreasing) we deduce

$$\int_{u\leq 4} u^2\, d\nu \;\leq\; 4 \int_{u\leq 4} h(u)\, d\nu.$$

Combining this with the nonnegativity of h and (3.1), we find

$$\int_{u\leq 4} \varphi\, |u|\, d\nu \;\leq\; 2 \left(\int_X \varphi^2\, d\nu\right)^{1/2} \left(\int_X h(u)\, d\nu\right)^{1/2} = 2 \left(\int_X \varphi^2\, d\nu\right)^{1/2} H(\mu|\nu)^{1/2}. \tag{3.3}$$

Since the function $t \longmapsto e^{2\sqrt{t}}$ is increasing and convex on $[1/4,+\infty)$ we can write

$$\exp\left(2\sqrt{\int_X \varphi^2\,d\nu}\right) \le \exp\left(2\sqrt{\int_X (\varphi+1/2)^2\,d\nu}\right) \le \int_X \exp\left(2\sqrt{(\varphi+1/2)^2}\right)\,d\nu$$
$$= \int_X e^{2\varphi+1}\,d\nu.$$

In other words,

$$2\sqrt{\int_X \varphi^2\,d\nu} \;\le\; 1 \,+\, \log\int e^{2\varphi}\,d\nu;$$

if we plug this into (3.3), we conclude that

$$\int_{u\le 4} \varphi\,|u|\,d\nu \;\le\; \left(1 \,+\, \log\int_X e^{2\varphi}\,d\nu\right)\,H(\mu|\nu)^{1/2}. \tag{3.4}$$

We now turn to the estimate of the **second term** ($u>4$) in (3.2). By applying the Young-type inequality

$$w\xi \;\le\; w\log w - w + e^{\xi} \qquad (w\ge 0, \xi\in\mathbb{R}) \tag{3.5}$$

with $w=u(x)$ and $\xi=\varphi(x)-Z$, where Z is a nonnegative constant to be chosen later, we find

$$\begin{aligned} u(x)\varphi(x) \;&\le\; u(x)\log u(x) \,-\, u(x) \,+ e^{\varphi(x)-Z} \,+ Zu(x) \\ &\le h(u(x)) \,+\, \left(\inf_{v>4}\sqrt{h(v)}\right)^{-1} e^{\varphi(x)-Z}\sqrt{h(u(x))} \,+ Zu(x) \end{aligned}$$

on $\{u(x)>4\}$. By integration, we deduce

$$\int_{u>4} u\varphi\,d\nu \;\le\; \int_{u>4} h(u)\,d\nu + \sqrt{k}\int_{u>4} e^{\varphi-Z}\sqrt{h(u)}\,d\nu + Z\int_{u>4} u\,d\nu,$$

where

$$k := \left(\inf_{v>4} h(v)\right)^{-1} = \frac{1}{h(4)} < \frac{1}{4}\cdot$$

By Cauchy-Schwarz inequality again,

$$\int_{u>4} e^{\varphi-Z}\,\sqrt{h(u)}\,d\nu \;\le\; \sqrt{\int_X e^{2(\varphi-Z)}\,d\nu}\sqrt{\int_{u>4} h(u)\,d\nu} = \sqrt{\int_X e^{2(\varphi-Z)}\,d\nu}\,\sqrt{H(\mu|\nu)}.$$

Finally, from the inequality

$$v\ge 4 \quad\Longrightarrow\quad v\le 4k\;h(v)$$

we deduce

$$\int_{u>4} u\,d\nu \;\le\; 4k\int_{u>4} h(u)\,d\nu \le 4kH(\mu|\nu).$$

Our conclusion is that, for any constant $Z \geq 0$,

$$\int_{u>4} \varphi\, u\, d\nu \;\leq\; (1+4kZ)H(\mu|\nu) \;+\; \sqrt{k}\sqrt{\int_X e^{2(\varphi-Z)}\, d\nu}\,\sqrt{H(\mu|\nu)}. \tag{3.6}$$

We now choose Z in such a way that

$$\int_X e^{2(\varphi-Z)}\, d\nu \;=\; 1;$$

in other words,

$$Z \;:=\; \frac{1}{2}\log\int e^{2\varphi}\, d\nu \geq 0.$$

Plugging this into (3.6), we conclude that

$$\int_{u>4} \varphi\, u\, d\nu \;\leq\; \left(1+2k\log\int e^{2\varphi}\, d\nu\right) H(\mu|\nu) + \;\sqrt{k}\,\sqrt{H(\mu|\nu)}. \tag{3.7}$$

Now inequality (i) follows from (3.4) and (3.7) upon noting that $1+\sqrt{k} < \dfrac{3}{2}$ and $2k < \dfrac{1}{2}\cdot$

We next turn to the **proof of (ii)**. Although the decomposition (3.2) and the same kind of argument would also lead to the result, we prefer to proceed as follows.

Since $h(0) = h'(0) = 0$, by Taylor's formula with integral remainder, we can write

$$h(u) \;=\; u^2\int_0^1 \frac{1-t}{1+tu}\, dt,$$

and thus

$$H(\mu|\nu) \;=\; \int_X\int_0^1 \frac{u^2(x)\,(1-t)}{1+tu(x)}\, d\nu(x)\, dt.$$

On the other hand, by Cauchy-Schwarz inequality on $(0,1)\times X$

$$\begin{aligned}\left(\int_0^1 (1-t)\, dt\right)^2 \left(\int_X \varphi|u|\, d\nu\right)^2 &= \left(\int_X\int_0^1 (1-t)\varphi|u|\, d\nu\, dt\right)^2 \\ &\leq \left[\int_X\int_0^1 (1-t)\,(1+tu)\,\varphi^2\, d\nu\, dt\right]\left[\int_x\int_0^1 \frac{1-t}{1+tu}|u|^2\, d\nu\, dt\right];\end{aligned}$$

thus

$$\left(\int_X \varphi\,|u|\, d\nu\right)^2 \;\leq\; CH(\mu|\nu)$$

where

$$C \;:=\; \frac{\displaystyle\iint (1-t)\,(1+tu)\,\varphi^2\, d\nu\, dt}{\left(\displaystyle\int_0^1 (1-t)\, dt\right)^2}\cdot \tag{3.8}$$

We decompose the numerator as follows:

$$\iint (1-t)\,(1+tu)\,\varphi^2\,d\nu\,dt = \int (1-t)t\,dt \int (1+u)\,\varphi^2\,d\nu \;+\; \int (1-t)^2\,dt \int \varphi^2\,d\nu$$
$$= \frac{1}{6}\int \varphi^2\,d\mu \,+\, \frac{1}{3}\int \varphi^2\,d\nu. \tag{3.9}$$

From the convexity inequality

$$\int \varphi^2\,d\mu \;\le\; H(\mu|\nu) + \log \int e^{\varphi^2}\,d\nu, \tag{3.10}$$

(a well-known consequence of (3.5), see for instance [6, eq. (5.13)]) and Jensen's inequality, in the form

$$\int \varphi^2\,d\nu \;\le\; \log \int e^{\varphi^2}\,d\nu, \tag{3.11}$$

we deduce that the right-hand side of (3.9) is bounded above by

$$\frac{1}{6}H(\mu|\nu) + \frac{1}{2}\log \int e^{\varphi^2}\,d\nu.$$

Plugging this into (3.8), we conclude that

$$\left(\int \varphi\,|u|\,d\nu\right)^2 \;\le\; \left(\frac{2}{3}H + 2L\right) H, \tag{3.12}$$

where H stands for $H(\mu|\nu)$ and L for $\log \displaystyle\int e^{\varphi^2}\,d\nu$.

The preceding bound is good only for "small" values of H. We now complement it with another bound which is relevant for "large" values of H. To do so, 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 (3.10) and (3.11).

Combining this with (3.12), 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

$$\min(at^2+bt,\; t+d) \;\le\; Mt, \qquad M \;=\; \frac{1}{2}\left\{1+b+\sqrt{(b-1)^2+4ad}\right\}$$

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. □

Remark 12. If $\varphi \equiv 1$, we can replace the inequality (3.10) by just $\int_x d\mu = 1$; then the first part of the proof of (ii) becomes a proof of the usual CKP inequality, with the sharp constant $\sqrt{2}$.

4. Application to random dynamical systems

Let now be given a Polish space X, an arbitrary element $x_0 \in X$ and a set of conditional Borel probability measures $(P_k(\cdot|x^{k-1}))_{x^{k-1}\in X^{k-1},\, k\ge 1}$, depending on $x^{k-1} = (x_1, \dots, x_{k-1}) \in X^{k-1}$ in a measurable way. We interpret x_0 as the (deterministic) initial position of a random dynamical system $(X_k)_{k\in\mathbb{N}}$, with values in X, and $P_k(\;\cdot\;|x^{k-1})$ as the law of X_k, knowing that $X_0 = x_0$ and $(X_1, \dots, X_{k-1}) = x^{k-1}$. The question is whether it is possible, knowing some nice bounds on the conditional probability measures, to get a T_1 inequality for the law P^n of $(X_1, \dots, X_n)$ on X^n, with a nice dependence on n.

Let us first assume that all the conditional probability measures satisfy a T_1 inequality, say with a uniform constant. In the context of independent random variables, it is rather easy [6, p. 122] to show that P^n satisfies $T_1(\lambda)$ for $\lambda^{-1} = O(n)$, and that this is sharp in general. Now we want to know whether the same behavior is generic for dependent random variables. Some results in that direction have been obtained by Marton and by Rio; they are summarized and slightly improved in [5]. In those references it is shown that if each $P_k(\;\cdot\;|x^{k-1})$ satisfies $T_1(\kappa)$ for some fixed $\kappa > 0$, and the random dynamical system is weakly dependent, in the sense that the future does not depend too much on the present, then the answer is positive. See [5, Section 4] for precise assumptions. For instance, a sufficient condition is that the dynamical system is Markovian and that the map

$$x_{k-1} \longmapsto P_k(\;\cdot\;|x^{k-1})$$

is L-Lipschitz from X to $P(X)$, equipped with the W_1 distance, uniformly in k, for some $L < 1$.

In the present section, we shall establish a variant of this result under a different set of assumptions, which seems to be easier to check in practical situations, because it is expressed in terms of exponential moments with respect to a given origin point (which we chose, arbitrarily, as the starting point of the dynamical system). What will make our argument work (in a very straightforward way) is the simple and explicit dependence of the constants in Theorem 1 upon n when X is replaced by X^n.

In the sequel we consider a Polish space X, equipped with a measurable distance d, x_0 an arbitrary element in X, and $(P_k(\;\cdot\;|x^{k-1}))_{x^{k-1}\in X^{k-1},\, k\ge 1}$ a family of Borel probability measures on X, depending on $x^{k-1} := (x_1, \dots, x_{k-1}) \in X^{k-1}$ in a measurable way. For all $n \geq 1$, we define the probability measure P^n on X^n by

$$dP^n(x_1, \dots, x_n) = dP_1(x_1)\, dP_2(x_2|x_1)\, \dots\, dP_n(x_n|x_1, \dots, x_{n-1}),$$

and equip X^n with the distance D defined by

$$D(x,y) = D_2(x,y) := \sqrt{\sum_{k=1}^n d(x_k, y_k)^2}.$$

There is an important difference with the above-mentioned works, namely the choice of the distance on the product space X^n: instead of D_2, they consider the distance

$$D_1(x,y) := \sum_{k=1}^n d(x_k, y_k).$$

While D_2 is often more natural than D_1, the latter is better adapted for arguments involving tensorization and Lipschitz functions. Of course, $D_1 \leq \sqrt{n} D_2$, so the distance D_2 is stronger than D_1 for each finite n, but does not behave similarly in the asymptotic regime $n \to +\infty$. Accordingly, if we try to deduce natural concentration estimates from our results, we typically obtain

$$P^n \left[\left| \frac{1}{n} \sum_{k=1}^n \varphi(x_k) - \int \left(\frac{1}{n} \sum_{k=1}^n \varphi(x_k) \right) dP^n(x^n) \right| \geq \varepsilon \right] \leq 2 \exp\left(-\frac{\lambda \varepsilon^2}{2} \right)$$

for any $n \geq 1$ and any Lipschitz function φ on X with Lipschitz seminorm 1. The fact that this bound does not go to 0 as $n \to \infty$ is probably linked to Remark 2 (2).

Theorem 13 (T_1 inequalities for random dynamical systems)**.** *With the above notation, assume the existence of $\alpha_0 > 0$, a sequence $(z_k)_{k \geq 1}$ in X and families of nonnegative numbers $(\gamma_k)_{k \geq 1}$, $(\beta_j)_{j \geq 1}$ with*

$$\gamma := \sup_{n \geq 1} \left[\frac{1}{n} \sum_{k=1}^n \gamma_k \right] < +\infty, \qquad \beta := \sum_{j \geq 1} \beta_j < \alpha_0,$$

such that for all $k \geq 1$, $x^{k-1} \in X^{k-1}$,

$$\log \int_X e^{\alpha_0\, d(z_k, x_k)^2}\, dP_k(x_k|x^{k-1}) \leq \gamma_k + \sum_{j=1}^{k-1} \beta_j\, d(z_{k-j}, x_{k-j})^2.$$

Then, there exists $\lambda > 0$ such that for all $n \geq 1$, P^n satisfies $T_1(\lambda/n)$.

Particular case 14. Consider a homogeneous Markov chain on X with transition kernel $P(dy|x)$. Assume the existence of $(x_0, y_0) \in X \times X$, $\alpha_0 > 0$, $\beta < \alpha_0$ and $C < +\infty$ such that

$$\forall x \in X, \qquad \int_X e^{\alpha_0\, d(y_0, y)^2}\, P(dy|x) \leq C e^{\beta\, d(x_0, x)^2}. \tag{4.1}$$

Then there exists $\lambda > 0$ such that for all $n \geq 1$, P^n satisfies $T_1(\lambda/n)$.

Remark 15. If Condition (4.1) is satisfied for some choice of $(x_0, y_0, \alpha_0, \beta, C)$, then for any $\alpha_0' < \alpha_0$ and $(x_0', y_0') \in X \times X$ we can find $\beta' \in [\beta, \alpha_0')$, $C' < +\infty$ such that Condition (4.1) is satisfied for $(x_0', y_0', \alpha_0', \beta', C')$. Thus the choice of reference points x_0 and y_0 is arbitrary: for instance, if $X = \mathbb{R}^d$, we can choose 0 for both, and the condition becomes

$$\exists \alpha > 0,\ \beta < \alpha,\ C < +\infty; \qquad \forall x \in \mathbb{R}^d, \qquad \int_{\mathbb{R}^d} e^{\alpha |y|^2}\, P(dy|x) \le C e^{\beta |x|^2}. \tag{4.2}$$

Proof of Theorem 13. Let $\alpha := \alpha_0 - \beta$. Since $\alpha \le \alpha_0$, by assumption,

$$\int_X e^{\alpha d(x_0, x_n)^2}\, P_n(dx_n | x^{n-1}) \le \exp\left(\gamma_n + \sum_{k=1}^{n-1} \beta_k d(z_{n-k}, x_{n-k})^2 \right).$$

In particular,

$$\int_{X^n} e^{\alpha D(z^n, x^n)^2}\, P^n(dx^n) \le e^{\gamma_n} \int_{X^n} \exp\left(\sum_{k=1}^{n-1} (\alpha + \beta_k) d(z_{n-k}, x_{n-k})^2 \right) P^{n-1}(dx^{n-1}).$$

Here $z^n = (z_1, \ldots, z_n)$; note that $\alpha + \beta_k \le \alpha_0$ for all k, and in particular we can repeat the argument with $n-1$ in place of n. Using an induction argument, one easily shows that

$$\int_{X^n} e^{\alpha D(x^n, z^n)^2}\, P^n(dx^n) \le e^{\sum_{k=1}^n \gamma_k} \le e^{n\gamma}.$$

In particular,

$$\log \int_{X^n} e^{\alpha D(x^n, z^n)^2}\, P^n(dx^n) = O(n),$$

and we conclude by applying the results presented in section 2. □

As examples of application we now consider the following two particular cases:

Example 16. Let (X_i) be a Markovian dynamical system on a Polish space X, with transition kernel $P(\,\cdot\,|x)$ such that

(i) $P(\,\cdot\,|x)$ satisfies $T_1(\lambda)$ for a constant λ independent of x;
(ii) the map $x \mapsto P(\,\cdot\,|x)$ is L-Lipschitz from X to $P(X)$, equipped with the W_1 distance, with $L < 1$.

Then there exist $\alpha > 0$ and $\beta < \alpha$ such that for any $x_0, y_0 \in X$, there exists $\gamma < +\infty$ such that

$$\log \int_X e^{\alpha\, d(y_0, y)^2}\, P(dy|x) \le \gamma + \beta\, d(x_0, x)^2$$

for all $x \in X$. In particular the hypotheses of Theorem 13 hold in view of the Particular case 14.

Example 17. Let $(X_k)_{k \in \mathbb{N}}$ be a dynamical system on $\mathbb{R}^d$ such that the hypotheses of [5, Theorem 4.1] hold, that is, with the notation introduced above,

(i) there exists some constant λ such that

$$W_1(\nu, P_k(\,\cdot\,|x^{k-1})) \;\leq\; \sqrt{\frac{2}{\lambda} H\big(\nu | P_k(\,\cdot\,|x^{k-1})\big)}$$

for all $k \geq 1$, x^{k-1} in $(\mathbb{R}^d)^{k-1}$ and all probability measures ν on $\mathbb{R}^d$;

(ii) there exist some nonnegative numbers a_j such that $\displaystyle\sum_{j=1}^{+\infty} a_j < 1$ and

$$W_1\big(P_k(\,\cdot\,|x^{k-1}), P_k(\,\cdot\,|\tilde{x}^{k-1})\big) \;<\; \sum_{j=1}^{k-1} a_j |x_{k-j} - \tilde{x}_{k-j}|$$

for all $k \geq 1$ and $x^{k-1}, \tilde{x}^{k-1}$ in $(\mathbb{R}^d)^{k-1}$.

Then the assumptions of Theorem 13 also hold for this system.

This last example shows that our assumptions are not less general than those in [5]. Note carefully that when we apply Theorem 13 to this system, we do not recover such a strong conclusion as in [5] because of the choice of distances on product spaces (D_2 instead of D_1).

Since the proofs for both Examples 16 and 17 are similar, we only study the second example.

Proof of the assertion in Example 17. In a **first step** we prove that for any $k \geq 1$, x^{k-1}, z^{k-1} in $(\mathbb{R}^d)^{k-1}$, z_k in $\mathbb{R}^d$, $\varepsilon, \delta > 0$ and $a < \dfrac{\lambda}{2}$, we have

$$\begin{aligned}\log \int & e^{a(1-\varepsilon)|y_k - z_k|^2} \, P_k(dy_k|z^{k-1}) \\ & \leq -\frac{1}{2}\log(1 - \frac{2a}{\lambda}) + a\left(\frac{1}{\varepsilon} - 1\right)\left(1 + \frac{1}{\delta}\right)\left(\int |t_k - z_k| \, P_k(dt_k|z^{k-1})\right)^2 \\ & \qquad + a\left(\frac{1}{\varepsilon} - 1\right)(1+\delta) \sum_{j=1}^{k-1} a_{k-j} \sum_{j=1}^{k-1} a_{k-j}|x_j - z_j|^2.\end{aligned}$$

Indeed, the probability measure $P_k(\,\cdot\,|x^{k-1})$ satisfies $T_1(\lambda)$ and the map $y \mapsto |y - z_k|$ is 1-Lipschitz, so by the Bobkov-Götze formulation of the T_1 inequality (see [3, Theorem 1.3] and [5, Section 1]) we have

$$\int e^{a\left[|y_k - z_k| - \int |t_k - z_k| \, P_k(dt_k|x^{k-1})\right]^2} P_k(dy_k|x^{k-1}) \;\leq\; \frac{1}{\sqrt{1 - 2a/\lambda}} \tag{4.3}$$

for any $a < \dfrac{\lambda}{2}$, $z_k \in \mathbb{R}^d$ and $x^{k-1} \in (\mathbb{R}^d)^{k-1}$.

Let then ε be some positive number. Integrating the inequality

$$(1-\varepsilon)|y_k - z_k|^2 \leq \left| |y_k - z_k| - \int |t_k - z_k| \, P_k(dt_k|x^{k-1}) \right|^2 + \left(\frac{1}{\varepsilon} - 1\right)\left(\int |t_k - z_k| \, P_k(dt_k|x^{k-1})\right)^2$$

and using (4.3) lead to

$$\log \int e^{a(1-\varepsilon)|y_k-z_k|^2}\, P_k(dy_k|x^{k-1}) \;\le\; -\frac{1}{2}\log(1-\frac{2a}{\lambda}) + a\left(\frac{1}{\varepsilon}-1\right)\left(\int |t_k-z_k|\, P_k(dt_k|x^{k-1})\right)^2 .$$

Recall the Kantorovich-Rubinstein formulation of the W_1 distance [11, Theorem 1.14]:

$$W_1(\mu,\nu) = \sup_{g\ 1-\mathrm{Lipschitz}} \left(\int g\, d\mu - \int g\, d\nu\right)$$

This and Assumption (ii), with $\tilde{x}^{n-1}=z^{n-1}$, imply

$$\begin{aligned}\int |t_k-z_k|\, P_k(dt_k|x^{k-1}) - \int |t_k-z_k|\, P_k(dt_k|z^{k-1}) \;&\le\; W_1(P_k(\,\cdot\,,x^{k-1}),P_k(\,\cdot\,,z^{k-1}))\\ &\le\; \sum_{j=1}^{k-1} a_{k-j}\,|x_j-z_j|.\end{aligned}$$

Thus for any positive number δ

$$\begin{aligned}&\left(\int |t_k-z_k|\, P_k(dt_k|x^{k-1})\right)^2\\ &\qquad\le \left(1+\frac{1}{\delta}\right)\left(\int |t_k-z_k|\, P_k(dt_k|z^{k-1})\right)^2 + (1+\delta)\left(\sum_{j=1}^{k-1} a_{k-j}\,|x_j-z_j|\right)^2,\end{aligned}$$

and by Cauchy-Schwarz inequality we can bound this quantity by

$$\left(1+\frac{1}{\delta}\right)\left(\int |t_k-z_k|\, P_k(dt_k|z^{k-1})\right)^2 + (1+\delta)\sum_{j=1}^{k-1} a_{k-j}\sum_{j=1}^{k-1} a_{k-j}\,|x_j-z_j|^2.$$

This concludes this first step.

In the **second step**, we build the sequence (z_k) by the following induction process. Let z_1 be arbitrary in $\mathbb{R}^d$; assuming that we have defined $z^{k-1}=(z_1,\dots,z_{k-1})$, we let

$$z_k \;:=\; \int_{\mathbb{R}^d} t_k\, P_k(dt_k|z^{k-1}).$$

Then

$$\begin{aligned}a\left(\int |t_k-z_k|\, P_k(dt_k|z^{k-1})\right)^2 &\le \log\int e^{a\,|t_k-z_k|^2}\, P_k(dt_k|z^{k-1})\\ &= \log\int e^{a\left|t_k-\int t_k\, P_k(dt_k|z^{k-1})\right|^2}\, P_k(dt_k|z^{k-1})\\ &\le \log\frac{1}{\sqrt{1-2a/\lambda}}\end{aligned}$$

thanks to Jensen's inequality and again the Bobkov-Götze formulation of the T_1 inequality, which is satisfied by $P_k(\,\cdot\,|z^{k-1})$.

Now we choose $\varepsilon \in (0,1)$ and $\delta > 0$ in such a way that

$$a\left(\frac{1}{\varepsilon}-1\right)(1+\delta)\sum_{j=1}^{+\infty} a_j \, < a(1-\varepsilon) :$$

for instance, $\varepsilon := \left(\sum_{j=1}^{+\infty} a_j\right)^{1/2}$ and $\delta := \left(\sum_{j=1}^{+\infty} a_j\right)^{-1/4} - 1$ will do. Then the assumptions of Theorem 13 can be checked to hold for

$$\alpha_0 := a(1-\varepsilon),$$

$$\gamma_k := -\frac{1}{2}\log\left(1-\frac{2a}{\lambda}\right)\left[1+\left(\frac{1}{\varepsilon}-1\right)\left(1+\frac{1}{\delta}\right)\right]$$

and

$$\beta_j := a\left(\frac{1}{\varepsilon}-1\right)(1+\frac{1}{\delta})\, a_j.$$

□

Acknowledgments: Both authors thank Patrick Cattiaux and Arnaud Guillin for stimulating discussions and their constant interest through the elaboration of this work, which as a matter of fact was triggered by a collaboration with Guillin. An anonymous referee is thanked for his relevant comments which helped improve the presentation of the paper.

References

[1] Blower, G. The Gaussian isoperimetric inequality and transportation. *Positivity. 7*, (2003), 203–224.

[2] Bobkov, S., Gentil, I., and Ledoux, M. Hypercontractivity of Hamilton-Jacobi equations. *J. Math. Pures Appl. 80*, 7 (2001), 669–696.

[3] Bobkov, S., and Götze, F. Exponential integrability and transportation cost related to logarithmic Sobolev inequalities. *J. Funct. Anal. 163* (1999), 1–28.

[4] Cattiaux, P., and Guillin, A. A criterion for Talagrand's quadratic transportation cost. Preprint, 2004. Available at `http://arxiv.org/abs/math.PR/0312081`

[5] Djellout, H., Guillin, A., and Wu, L. Transportation cost-information inequalities for random dynamical systems and diffusions. *Ann. Probab.32*, (2004), 2702–2732.

[6] Ledoux, M. *The concentration of measure phenomenon*, vol. 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2001.

[7] Marton, K. Bounding $\overline{d}$-distance by information divergence: a method to prove measure concentration. *Ann. Probab. 24* (1996), 857–866.

[8] Otto, F., and Villani, C. Generalization of an inequality by Talagrand, and links with the logarithmic Sobolev inequality. *J. Funct. Anal. 173* (2000), 361–400.

[9] Rachev, S. T. *Probability metrics and the stability of stochastic models.* John Wiley & Sons Ltd., Chichester, 1991.

[10] Talagrand, M. Transportation cost for Gaussian and other product measures. *Geom. Funct. Anal. 6*, no.3 (1996), 587–600.

[11] Villani, C. *Topics in optimal transportation*, vol. 58 of *Graduate Studies in Mathematics.* American Mathematical Society, Providence, RI, 2003.