Continuous (convex) optimisation

M2 - PSL / Dauphine / S.U.

Antonin Chambolle, CNRS, CEREMADE

Université Paris Dauphine PSL

Oct.-Dec. 2023

Lecture 6: Non-linear problems, mirror descent.

Contents

Nonlinear norms

Most of the time we will work in finite dimension. However the general setting we can consider here is of a Banach space $\mathcal{X}$ with dual $\mathcal{X}^*$ and respective norms denoted $\|\cdot\|$, $\|\cdot\|_*$ with

$$\|y\|_* = \sup\{\langle y, x\rangle_{\mathcal{X}^*,\mathcal{X}} : \|x\| \le 1\} \qquad \|x\| = \sup\{\langle y, x\rangle_{\mathcal{X}^*,\mathcal{X}} : \|y\|_* \le 1\}.$$

Now, given f a C^1 function, one can define its differential:

$$f(x') = f(x) + \langle df(x), x' - x\rangle_{\mathcal{X}^*,\mathcal{X}} + o(\|x' - x\|)$$

but there is no obvious notion of a "Gradient".

Nonlinear Gradient descent

However, we can easily generalize the gradient descent as follows: given x^k, we let x^{k+1} be a minimizer of

$$\min_x f(x^k) + \left\langle df(x^k), x - x^k \right\rangle_{\mathcal{X}^*,\mathcal{X}} + \frac{1}{2\tau}\|x - x^k\|^2$$

provided such a minimizer exists. This will be the case for instance

- In finite dimension;
- If $\mathcal{X}$ is reflexive (or if $\mathcal{X}$ is a dual and f is weakly-$*$ lsc).

We assume one of these conditions hold.

Nonlinear Gradient descent

Convergence

As in the linear case, we can show the following:

Theorem

Assume df L-Lipschitz and consider the iterates x^k of the non-linear gradient descent with $\tau = 1/L$. Then, if x^* is a minimizer and $C = \max_{\{f(x) < f(x^0)\}} \|x - x^*\|^2 < +\infty$, one has the rate:

$$f(x^k) - f(x^*) \leq \frac{2LC}{k+1}.$$

Functions with Lipschitz differential

Of course, we say that f is a function with Lipschitz differential $df(x)$ iff for any $x, x' \in \mathcal{X}$,

$$\|df(x) - df(x')\|_* \leq \|x - x'\|$$

where each norm has to be taken in the appropriate space. Then, one has, exactly as before,

$$\begin{aligned}
f(x') &= f(x) + \int_0^1 \langle df(x + s(x' - x)), x' - x\rangle \, ds \\
&= f(x) + \langle df(x), x' - x\rangle + \int_0^1 \langle df(x + s(x' - x)) - df(x), x' - x\rangle \, ds \\
&\leq f(x) + \langle df(x), x' - x\rangle + \int_0^1 \|df(x + s(x' - x)) - df(x)\|_* \|x' - x\| ds \\
&\leq f(x) + \langle df(x), x' - x\rangle + \int_0^1 Ls\|x' - x\|^2 ds = f(x) + \langle df(x), x' - x\rangle + \frac{L}{2}\|x' - x\|^2.
\end{aligned}$$

Dual norms

We show the following lemma:

Lemma

Let $\mathcal{F}(x) = \mu\|x\|^2/2$. Then its conjugate is $\mathcal{F}^*(y) = \|y\|_*^2/(2\mu)$.

Proof: we write

$$\mathcal{F}^*(y) = \sup_x \langle y, x\rangle - \frac{\mu}{2}\|x\|^2 = \sup_{t>0} \sup_{\|x\|\le t} \langle y, x\rangle - \frac{\mu t^2}{2} = \sup_{t>0} t\|y\|_* - \frac{\mu t^2}{2} = \frac{1}{2\mu}\|y\|_*^2.$$

□

Dual norms

We show the following lemma:

Lemma

Let $\mathcal{F}(x) = \mu\|x\|^2/2$. Then its conjugate is $\mathcal{F}^*(y) = \|y\|_*^2/(2\mu)$.

Proof: we write

$$\mathcal{F}^*(y) = \sup_x \langle y, x\rangle - \frac{\mu}{2}\|x\|^2 = \sup_{t>0} \sup_{\|x\|\le t} \langle y, x\rangle - \frac{\mu t^2}{2} = \sup_{t>0} t\|y\|_* - \frac{\mu t^2}{2} = \frac{1}{2\mu}\|y\|_*^2.$$

□

Legendre-Fenchel identity shows again that $y \in \partial\mathcal{F}(x) \Leftrightarrow x \in \partial\mathcal{F}^*(y) \Leftrightarrow \langle y, x\rangle = \mathcal{F}(x) + \mathcal{F}^*(y)$, yet in addition, being $\mathcal{F}$ and $\mathcal{F}^*$ positively 2-homogeneous, we have also $\langle y, x\rangle = 2\mathcal{F}(x) = 2\mathcal{F}^*(y)$ and $\mathcal{F}(x) = \mathcal{F}^*(y)$.

Nonlinear Gradient descent

Returning to the gradient descent algorithm, we have, since x^{k+1} is a minimizer (for $\mu = 1/\tau$) of:

$$\min_x f(x^k) + \left\langle df(x^k), x - x^k \right\rangle_{\mathcal{X}^*,\mathcal{X}} + \mathcal{F}(x - x^k) = f(x^k) - \mathcal{F}^*(-df(x^k)),$$

and $-df(x^k) \in \partial\mathcal{F}(x^{k+1} - x^k)$, $x^{k+1} - x^k \in -\partial\mathcal{F}^*(df(x^k))$, while $-\left\langle df(x^k), x^{k+1} - x^k \right\rangle_{\mathcal{X}^*,\mathcal{X}} = \mathcal{F}(x^{k+1} - x^k) + \mathcal{F}^*(-df(x^k))$.
In particular, the algorithm is defined by:

$$x^{k+1} = x^k - \tau p^k, \quad p^k \in \|df(x^k)\|_* \partial\|\cdot\|_*(df(x^k)).$$

By 2-homogeneity of $\mathcal{F}$ and $\mathcal{F}^*$ one also sees that $\mathcal{F}(x^{k+1} - x^k) = \mathcal{F}^*(-df(x^k))$.

Nonlinear Gradient descent

Now we return to the proof of a rate: In addition, since df is L-Lipschitz,

$$f(x^{k+1}) \leq f(x^k) + \left\langle df(x^k), x^{k+1} - x^k \right\rangle_{\mathcal{X}^*,\mathcal{X}} + \frac{L}{2}\|x^{k+1} - x^k\|^2$$

$$= f(x^k) + \frac{L}{2}\|x^{k+1} - x^k\|^2 - \mathcal{F}(x^{k+1} - x^k) - \mathcal{F}^*(-df(x^k)) = f(x^k) + \left(\frac{L}{2} - \frac{1}{2\tau}\right)\|x^{k+1} - x^k\|^2 - \frac{\tau}{2}\|df(x^k)\|_*^2.$$

so that if $\tau = 1/L$, one obtains:

$$f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}\|df(x^k)\|_*^2.$$

Now we can proceed as in the Euclidean setting. We observe that

$$f(x^*) \geq f(x^k) + \left\langle df(x^k), x^* - x^k \right\rangle \quad \Rightarrow \quad f(x^k) - f(x^*) \leq \|df(x^k)\|_* \|x^k - x^*\|$$

so that if $\Delta_k = f(x^k) - f(x^*)$,

$$\Delta_{k+1} \leq \Delta_k - \frac{\Delta_k^2}{2L\|x^k - x^*\|^2}.$$

Nonlinear Gradient descent: rate

If we now that $C_k = \max_{0\le i\le k} \|x^i - x^*\|^2 \le C$ remains bounded (for instance if $\{f \le f(x^0)\}$ is bounded) then we deduce:

$$f(x^k) - f(x^*) \le \frac{2LC}{k+1}$$

as in the Hilbertian case.

Strongly convex functions in non-Euclidean spaces

Are *not!!* functions f such that $f - \mu\|.\|^2/2$ is convex!

Definition

The function f is $\mu-$strongly convex if and only if for any $x, x' \in X$ and $t \in [0,1]$,

$$f(tx + (1-t)x') \leq tf(x) + (1-t)f(x') - \mu\frac{t(1-t)}{2}\|x - x'\|^2$$

Then one can show the following. We assume $\mathcal{X}$ is reflexive.

Theorem

Let $f : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ *be convex, proper, lower semi-continuous. Then* f *is strongly convex if and only if for all* $x, x' \in \mathcal{X}$ *and all* $y \in \partial f(x)$*, one has:*

$$f(x') \geq f(x) + \langle y, x' - x\rangle_{\mathcal{X}^*,\mathcal{X}} + \frac{\mu}{2}\|x - x'\|^2.$$

Strongly convex functions

Proof. One direction is easy (and does not require lower semicontinuity): if f is strongly convex and $x, x' \in \mathcal{X}$, $y \in \partial f(x)$, then for any $t \in (0,1)$,

$$f(tx + (1-t)x') \geq f(x) + (1-t)\left\langle y, x' - x \right\rangle.$$

From the strong convexity, we deduce

$$f(x) + (1-t)\left\langle y, x' - x \right\rangle \leq tf(x) + (1-t)f(x') - \mu\frac{t(1-t)}{2}\|x - x'\|^2.$$

Dividing by $(1-t)$ it follows:

$$f(x') \geq f(x) + \left\langle y, x' - x \right\rangle + \mu\frac{t}{2}\|x - x'\|^2$$

and letting $t \to 1$ we conclude.

Strongly convex functions

For the converse, we need to use points where the subgradient exists. Let $x, x' \in \mathcal{X}$ and $t \in [0,1]$. We do as follows: we let $x_t = tx + (1-t)x'$ and assume $f(x), f(x')$ are finite (otherwise, nothing to prove). Let ξ_n be a minimizer of:

$$\min_{\xi} f(\xi) + \frac{n}{2}\|\xi - x_t\|^2$$

Being $\|\cdot\|$ strongly continuous, one can show that a solution (which exists because a minimizing sequence is bounded, hence weakly converging since we assumed $\mathcal{X}$ is reflexive, and Hahn-Banach's theorem then shows that f is weakly lsc.) satisfies:

$$\partial f(\xi_n) + n\|\xi_n - x_t\|\partial\|\cdot\|(\xi_n - x_t) \ni 0 \quad \Leftrightarrow \quad \eta_n := -n\|\xi_n - x_t\|\partial\|\cdot\|(\xi_n - x_t) \in \partial f(\xi_n).$$

Using

$$f(\xi_n) + \frac{n}{2}\|\xi_n - x_t\|^2 \le f(x_t) \le tf(x) + (1-t)f(x') < +\infty,$$

we deduce that $\xi_n \to x_t$, then that $f(x_t) \le \liminf_n f(\xi_n)$, and eventually that

$$\frac{n}{2}\|\xi - x_t\|^2 \to 0$$

as $n \to \infty$.

Strongly convex functions

Now, we can write:

$$\begin{cases} f(x) \geq f(\xi_n) + \langle \eta_n, x - \xi_n \rangle + \frac{\mu}{2}\|x - \xi_n\|^2 \\ f(x') \geq f(\xi_n) + \langle \eta_n, x' - \xi_n \rangle + \frac{\mu}{2}\|x' - \xi_n\|^2. \end{cases}$$

We multiply the first equation by t and the second by $(1 - t)$, and sum:

$$tf(x) + (1 - t)f(x') \geq f(\xi_n) + \langle \eta_n, x_t - \xi_n \rangle + \frac{\mu}{2}(t\|x - \xi_n\|^2 + (1 - t)\|x' - \xi_n\|^2).$$

As $\|\cdot\|$ is positively 1-homogeneous, Euler's identity shows $\langle \eta_n, x_t - \xi_n \rangle = n\|x_t - \xi_n\|^2 \to 0$ as $n \to \infty$. In the limit (and because f is lsc) we find

$$tf(x) + (1 - t)f(x') \geq f(x_t) + \frac{\mu}{2}(t\|x - x_t\|^2 + (1 - t)\|x' - x_t\|^2)$$
$$= f(x_t) + \frac{\mu}{2}(t(1 - t)^2\|x - x'\|^2 + (1 - t)t^2\|x' - x\|^2) = f(x_t) + \mu\frac{t(1 - t)}{2}\|x - x'\|^2$$

□

Strongly convex functions and Lipschitz differentials

Now, we have the following theorem, which is a duality result between convex functions with Lipschitz differential and strongly convex functions:

Theorem

Let f be convex, lsc. Then f has (L-)Lipschitz differential if and only if f^* is ($1/L$-)strongly convex.

Proof: If f is convex with L-Lipschitz differential, then one has for all x, x'

$$f(x') \leq f(x) + \left\langle df(x), x' - x \right\rangle + \frac{L}{2}\|x - x'\|^2.$$

We let $y = df(x)$ so that, by Legendre-Fenchel's identity, $x \in \partial f^*(y)$ and $\langle y, x \rangle = f(x) + f^*(y)$. Taking the conjugate of the inequality at a point y', we have

$$f^*(y') \geq \sup_{x'} \left\langle y', x' \right\rangle - f(x) - \left\langle y, x' - x \right\rangle - \frac{L}{2}\|x - x'\|^2 = f^*(y) + \sup_{x'} \left\langle y' - y, x' \right\rangle - \frac{L}{2}\|x - x'\|^2.$$

Strongly convex functions and Lipschitz differentials

Now, we recall that

$$\left(\frac{L}{2}\|\cdot\|^2\right)^*(p) = \frac{1}{2L}\|p\|_*^2.$$

We deduce

$$\sup_{x'} \left\langle y' - y, x' \right\rangle - \frac{L}{2}\|x - x'\|^2 = \left\langle y' - y, x \right\rangle + \sup_{x'} \left\langle y' - y, x' - x \right\rangle - \frac{L}{2}\|x - x'\|^2 = \left\langle y' - y, x \right\rangle + \frac{1}{2L}\|y' - y\|_*^2,$$

so that

$$f^*(y') \geq f^*(y) + \left\langle y' - y, x \right\rangle + \frac{1}{2L}\|y' - y\|_*^2 \qquad (*)$$

so that f^* is $(1/L)$-convex. Conversely, if $y, y' \in \mathcal{X}^*$ and $x \in \partial f^*(y)$, the same computation will show that if $(*)$ holds: (using $y \in \partial f(x)$ and $\langle y, x \rangle = f(x) + f^*(y)$):

$$f(x') \leq f(x) + \left\langle y, x' - x \right\rangle + \frac{L}{2}\|x' - x\|^2.$$

Since $f(x') \geq f(x) + \langle y, x' - x \rangle$, we deduce that f is differentiable at x and $y = df(x)$.

Strongly convex functions and Lipschitz differentials

In addition, if $y' \in \partial f(x')$ (hence as above, $y' = df(x')$) so that $x' \in \partial f^*(y')$, we write:

$$f^*(y') \geq f^*(y) + \left\langle y' - y, x \right\rangle + \frac{1}{2L}\|y' - y\|_*^2 \text{ and } f^*(y) \geq f^*(y') + \left\langle y - y', x' \right\rangle + \frac{1}{2L}\|y - y'\|_*^2$$

and we deduce $\langle x' - x, y' - y \rangle \geq \|y - y'\|_*^2 / L$. Since $\langle x' - x, y' - y \rangle \leq \|x - x'\| \|y - y'\|_*$ it follows that $\|df(x) - df(x')\|_* = \|y - y'\|_* \leq L\|x - x'\|$.
It remains to check that df is defined everywhere. Observe that f is globally bounded by a quadratic function hence locally finite, hence locally Lipschitz. Then, if $x_n \to x$ are points where a subgradient (hence differential) exists, since $df(x_n)$ is a Cauchy sequence: there exists $y \in \mathcal{X}^*$ with $df(x_n) \to y$ and we pass to the limit in:

$$f(x') \geq f(x_n) + \left\langle df(x_n), x' - x_n \right\rangle$$

to conclude that $p \in \partial f(x)$ so that $y = df(x)$. Hence f is C^1 with Lipschitz gradient. □

Example

A typical example is given by the entropy in $\mathbb{R}^d$ on the unit simplex $\Sigma := \{x \in \mathbb{R}^d : x_i \geq 0, \sum_i x_i = 1\}$:

$$\xi(x) = \begin{cases} \sum_i x_i \log x_i & \text{if } x \in \Sigma \\ +\infty & \text{else} \end{cases}$$

(where $0 \log 0$ is defined as 0). Then, one shows that the conjugate is the "log-sum-exp" function:

$$\xi^*(y) = \log \sum_i \exp(y_i)$$

also called "soft-max" since $\varepsilon\xi^*(y/\varepsilon)$ is an approximation of the max as $\varepsilon \to 0$.

Example

Pinsker inequality

Then, one can show the following:

Lemma (Pinsker inequality)

ξ is 1-strongly convex in the ℓ^1 norm.

That is, for any $x, x' \in \Sigma$ the unit simplex, $p \in \partial\xi(x)$,

$$\xi(x') - \xi(x) - \langle p, x' - x\rangle = \sum_i x_i' \log \frac{x_i'}{x_i} \geq \frac{1}{2}\left(\sum_i |x_i - x_i'|\right)^2$$

This latter inequality is called the "Pinsker inequality".

Example

Proof: We leave as an exercise that if $\|\cdot\| = \|\cdot\|_1$ is the ℓ^1 norm, then $\|\cdot\|_* = \|\cdot\|_\infty$.

First we prove the expression for ξ^*: one has to compute $\sup_{x\in\Sigma}\sum_i x_i y_i - x_i\log x_i$. For the maximum x there is a Lagrange multiplier λ for the constraint $\sum_i x_i = 1$ and one has $y_i - \log x_i - 1 = \lambda$ (and in particular $\xi^*(x) = \sum_i x_i(\lambda+1) = \lambda + 1 =: \lambda'$). One has $x_i = \exp(y_i - \lambda')$ and since $\sum_i x_i = 1$, $\exp(-\lambda')\sum_i \exp(y_i) = 1$ so that $\exp(\lambda') = \sum_i \exp(y_i)$, and $\lambda' = \log\sum_i \exp(y_i)$.

Example

Now, we prove that ξ^* has 1-Lipschitz gradient. Observe that

$$
\begin{aligned}
\|d\xi^*(y') - d\xi^*(y)\|_1 &= \sup_{\|z\|_\infty \le 1} \left\langle z, d\xi^*(y') - d\xi^*(y) \right\rangle \\
&= \sup_{\|z\|_\infty \le 1} \left\langle z, \int_0^1 d^2\xi^*(y + s(y'-y)) \cdot (y'-y)ds \right\rangle \\
&\le \sup_{\|z(\cdot)\|_\infty \le 1} \int_0^1 \left\langle z(s), d^2\xi^*(y + s(y'-y)) \cdot \frac{y'-y}{\|y'-y\|_\infty} \right\rangle ds \|y'-y\|_\infty \\
&\le \int_0^1 L(y + s(y'-y)) ds \|y'-y\|_\infty
\end{aligned}
$$

where

$$
L(y) := \sup_{\sigma_i \in [-1,1], \tau_j \in [-1,1]} \sum_{i,j} \frac{\partial^2 \xi^*}{\partial y_i \partial y_j}(y) \sigma_i \tau_j .
$$

If we can show that $L(y) \le 1$ for all y, we are done.

Example

We now show that $L(y) \le 1$ for all $y \in \mathbb{R}^d$. First, letting (for a given $y \in \mathbb{R}^d$) $a_{i,j} := \partial^2_{i,j}\xi^*(y)$, we have

$$a_{i,j} = \theta_i \delta_{i,j} - \theta_i \theta_j$$

where $\theta_i = \exp(y_i)/\sum_k \exp(y_k)$ and $\delta_{i,j}$ is the Kronecker symbol. In particular, $\theta \in \Sigma$, and we see that $\sum_i a_{i,j} = 0$ for all j and $\sum_j a_{i,j} = 0$ for all i.

Then, let τ, σ be a maximizer. Let $\sigma'_i = 1$ if $\sum_j a_{i,j}\tau_j \ge 0$ and -1 else, and then $\tau'_j = 1$ if $\sum_i a_{i,j}\sigma'_i \ge 0$ and -1 else: one checks that (σ', τ') is also a maximizer. Hence one can restrict the maximisation problem over $\sigma_i, \tau_j \in \{-1, 1\}$ and in particular we see that

$$L(y) = \max_{\sigma_i \in \{-1,1\}} \sum_j \left| \sum_i a_{i,j}\sigma_i \right|.$$

Then, $\sum_i a_{i,j}\sigma_i = \sum_{i:\sigma_i=1} a_{i,j} - \sum_{i:\sigma_i=-1} a_{i,j} = 2\sum_{i:\sigma_i=1} a_{i,j}$ since $\sum_i a_{i,j} = 0$. Introducing the variable $\xi = 2\sigma - 1$, we find that the max is

$$\max_{\xi_i \in \{0,1\}} 2\sum_j \left| \sum_i \xi_i a_{i,j} \right|.$$

Example

Then, for all j,

$$\left|\sum_i \xi_i a_{i,j}\right| = \left|\xi_j\theta_j - (\xi\cdot\theta)\theta_j\right| = \theta_j\left|\xi_j - (\xi\cdot\theta)\right| = \begin{cases}\theta_j(1-\xi\cdot\theta) & \text{if } \xi_j = 1\\ \theta_j(\xi\cdot\theta) & \text{if } \xi_j = 0\end{cases}$$
$$= \xi_j\theta_j(1-\xi\cdot\theta) + (1-\xi_j)\theta_j(\xi\cdot\theta)$$

so that

$$\sum_j \left|\sum_i \xi_i a_{i,j}\right| = \xi\cdot\theta(1-\xi\cdot\theta) + (\xi\cdot\theta) - (\xi\cdot\theta)^2 = 2\xi\cdot\theta(1-\xi\cdot\theta).$$

We deduce

$$L(y) = 4\max_{\xi_i\in\{0,1\}}(\xi\cdot\theta)(1-\xi\cdot\theta) \le 4\max_{0\le t\le 1} t(1-t) = 1$$

□

Remark: we see that the max is reached for $\tau = \sigma$, minimizing $|\tau\cdot\theta| = |\sum_{\tau_i=1}\theta_i - \sum_{\tau_i=-1}\theta_i|$.

Bregman distances and Legendre function

We say a convex function ξ with domain $D \subset \mathcal{X}$ is "Legendre" (Rockafellar 1970, Chen-Teboulle 1993) if

(i) ξ is C^1 in the (relative) interior of D;

(ii) $\lim_{x \to \partial D} \|\nabla \xi(x)\| = +\infty$;

(iii) ξ is 1-convex.

In particular, $\partial \xi(x) = \emptyset$ for $x \in \partial D$, and, given f convex, lsc., then if x solves:

$$\min_x \xi(x) + f(x)$$

one must have $x \in \mathring{D}$ and $-\nabla \xi(x) \in \partial f(x)$

[If "relative" in (i) this needs to be adapted a bit)]

Bregman distances

Given ξ Legendre, we define for $x, x' \in \mathcal{X}$:

$$D_\xi(x', x) := \xi(x') - \xi(x) - \langle d\xi(x), x' - x \rangle$$

and we observe that $D_\xi(x', x) \geq 0$ (by convexity), moreover $D_\xi(x', x) \geq \|x' - x\|^2/2$ if (iii) holds.

One has the following result:

Lemma

Three-point inequality [Chen-Teboulle 1993, Tseng 2008] Let g be convex, lsc., and assume $\hat{x}$ is a minimiser of $\min_x D_\xi(x, \bar{x}) + g(x)$. Then for all x,

$$D_\xi(x, \bar{x}) + g(x) \geq D_\xi(\hat{x}, \bar{x}) + g(\hat{x}) + D_\xi(x, \hat{x}).$$

Bregman distances

Proof: one has by minimality that

$$d\xi(\hat{x}) - d\xi(\bar{x}) + \partial g(\hat{x}) \ni 0 \quad \Leftrightarrow \quad \partial g(\hat{x}) \ni d\xi(\bar{x}) - d\xi(\hat{x}).$$

Hence for all x,

$$g(x) \geq g(\hat{x}) + \langle d\xi(\bar{x}) - d\xi(\hat{x}), x - \hat{x} \rangle .$$

We deduce

$$\begin{aligned}
D_\xi(x,\bar{x}) + g(x) \geq & \xi(x) - \xi(\bar{x}) - \langle d\xi(\bar{x}), x - \bar{x} \rangle + g(\hat{x}) + \langle d\xi(\bar{x}) - d\xi(\hat{x}), x - \hat{x} \rangle \\
&= \xi(x) - \xi(\hat{x}) + \xi(\hat{x}) - \xi(\bar{x}) - \langle d\xi(\bar{x}), x - \hat{x} + \hat{x} - \bar{x} \rangle + g(\hat{x}) + \langle d\xi(\bar{x}) - d\xi(\hat{x}), x - \hat{x} \rangle \\
&= \xi(x) - \xi(\hat{x}) + \xi(\hat{x}) - \xi(\bar{x}) - \langle d\xi(\hat{x}), x - \hat{x} \rangle - \langle d\xi(\bar{x}), \hat{x} - \bar{x} \rangle + g(\hat{x}) \\
&= D_\xi(x,\hat{x}) + D_\xi(\hat{x},\bar{x}) + g(\hat{x}).
\end{aligned}$$

□

Mirror descent (explicit-implicit)

Let ξ be a Legendre function.
Assume the function f has L-Lipschitz gradient and g is such that one can compute for each k:

$$\min_{x \in \operatorname{dom} \xi} \frac{1}{\tau} D_\xi(x, x^k) + \left\langle df(x^k), x \right\rangle + g(x)$$

and let x^{k+1} be the solution. This is a "mirror-prox" algorithm. Then thanks to the "three points inequality" one can deduce the same as for the forward-backward descent: for any x, one has for τ small enough, letting $F = f + g$:

$$\frac{1}{\tau} D_\xi(x, x^k) + F(x) \geq F(x^{k+1}) + \frac{1}{\tau} D_\xi(x, x^{k+1})$$

Mirror descent

Thanks to:

$$\frac{1}{\tau} D_\xi(x, x^k) + F(x) \geq F(x^{k+1}) + \frac{1}{\tau} D_\xi(x, x^{k+1}) \qquad (*)$$

we deduce exactly as in the Euclidean case:

Convergence rate for the mirror descent

Assume there exists x^* a minimizer of F in $\operatorname{dom}\xi$. Then the mirror-prox algorithm produces a sequence which satisfies:

$$F(x^k) - F(x^*) \leq \frac{D_\xi(x^*, x^0)}{\tau k}.$$

As usual, we obtain this by taking $x = x^k$ and $x = x^*$ in the descent inequality $(*)$.

Mirror descent (explicit-implicit)

One has thanks to the 3-points inequality:

$$
\begin{aligned}
\frac{1}{\tau}D_\xi(x,x^k) + F(x) &\geq \frac{1}{\tau}D_\xi(x,x^k) + f(x^k) + \left\langle df(x^k), x - x^k \right\rangle + g(x) \\
&\geq \frac{1}{\tau}D_\xi(x^{k+1},x^k) + f(x^k) + \left\langle df(x^k), x^{k+1} - x^k \right\rangle + g(x^{k+1}) + \frac{1}{\tau}D_\xi(x,x^{k+1}).
\end{aligned}
$$

Now $f(x^k) + \left\langle df(x^k), x^{k+1} - x^k \right\rangle = f(x^{k+1}) - D_f(x^{k+1}, x^k)$ by definition so that:

$$
\frac{1}{\tau}D_\xi(x,x^k) + F(x) \geq \frac{1}{\tau}D_\xi(x^{k+1},x^k) - D_f(x^{k+1},x^k) + F(x^{k+1}) + \frac{1}{\tau}D_\xi(x,x^{k+1}).
$$

Now, if f has L-Lipschitz gradient then $D_f(x^{k+1},x^k) \leq L\|x^{k+1} - x^k\|^2/2$, while ξ being strongly convex, $D_\xi(x^{k+1},x^k) \geq \|x^{k+1} - x^k\|^2/2$. Hence one finds that if $\tau \leq 1/L$,

$$
\frac{1}{\tau}D_\xi(x^{k+1},x^k) - D_f(x^{k+1},x^k) \geq 0
$$

and this ends the proof.

Relative smoothness

However, here, we need the strong convexity of ξ and the Lipschitz gradient of f only to bound the difference $D_\xi(x^{k+1}, x^k)/\tau - D_f(x^{k+1}, x^k)$. So a much simper and better assumption could be "there exists L such that $LD_\xi - D_f \geq 0$". When is it true??? Observe that by construction,

$$D_{f-g} = D_f - D_g$$

so that clearly, $D_f \geq D_g$ for any points if and only if $f - g$ is convex. Hence:

Definition

One says that f is L-relatively smooth with respect to ξ if $L\xi - f$ is convex.

Corollary

The nonlinear forward-backward algorithm has the rate $O(1/k)$ (when a minimizer exists) as soon as f is L-relatively smooth wr. ξ and $\tau \leq 1/L$.

(No L-Lipschitz or strongly convexity assumption needed here → "NoLips" algorithm (Bauschke, Bolte, Teboulle 2017). Can be improved with over-relaxation which *depends* on ξ.)

Relative strong convexity

Similarly (Teboulle 2018, Lu, Freund, Nesterov 2018, C-Pock 2016):

Definition

One says that f is relatively strongly convex wr. ξ if there exists $\gamma > 0$ such that $f - \gamma\xi$ is convex.

In case f or g is relatively strongly convex, one obtains a linear convergence rate. Indeed, the three-points inequality is improved to:

$$D_\xi(x, \bar{x}) + g(x) \geq D_\xi(\hat{x}, \bar{x}) + g(\hat{x}) + (1 + \mu_g) D_\xi(x, \hat{x}),$$

and the descent inequality is improved as before to, for $\tau \leq 1/L$:

$$\frac{1 - \tau\mu_f}{\tau} D_\xi(x, x^k) + F(x) \geq F(x^{k+1}) + \frac{1 + \tau\mu_g}{\tau} D_\xi(x, x^{k+1})$$

Accelerated Mirror descent

[Nesterov, Tseng]

Unfortunately, there is no way to accelerate under the mere assumption of relative smoothness, nor can we improve easily this method when f is relatively strongly convex. (cf Dragomir, Taylor, D'Aspremont, Bolte 2019.)

Assuming ξ is 1-convex and ∇f is L-Lipschitz, on the other hand, makes acceleration is possible. This is improved in addition under a relative strong convexity assumption.

The "accelerated mirror descent" is a possibility, the "accelerated primal-dual" algorithm another. We now explain the mirror descent algorithm in the simplest case, that is non relatively strongly convex.

Accelerated Mirror descent

The general algorithm is as follows: we assume f is has L-Lipschitz gradient. Let also g such that $\min_x \alpha g(x) + \xi(x) + \langle p, x\rangle$ is easily computed.
We pick x^0, set $y^0 = z^0 = x^0$, let $\alpha_0 = \beta_0 = 0$.

1. Let α_{k+1} be the largest root of:
$$\beta_{k+1} := \beta_k + \alpha_{k+1} = L\alpha_{k+1}^2;$$
2. Let: $x^{k+1} = (\alpha_{k+1} z^k + \beta_k y^k)/\beta_{k+1}$
3. Define z^{k+1} as the minimizer of
$$\min_x \frac{1}{\alpha_{k+1}} D_\xi(z, z^k) + (g(z) + f(x^{k+1}) + \left\langle df(x^{k+1}), z - x^{k+1}\right\rangle$$
4. Let $y^{k+1} = (\alpha_{k+1} z^{k+1} + \beta_k y^k)/\beta_{k+1}$; return to 1.

Accelerated Mirror descent

We prove that, letting $F = f + g$:

Rate of convergence for accelerated mirror descent.

$$F(y^k) - F(x^*) \leq \frac{4L}{k^2} D_\xi(x^*, y^0).$$

Accelerated Mirror descent

We prove that, letting $F = f + g$:

Rate of convergence for accelerated mirror descent.

$$F(y^k) - F(x^*) \leq \frac{4L}{k^2} D_\xi(x^*, y^0).$$

Proof: As in the descent lemma, we have that

$$\alpha_{k+1}(f(z) + g(z)) + D_\xi(z, z^k) \geq \alpha_{k+1}(g(z) + f(x^{k+1}) + \left\langle df(x^{k+1}), z - x^{k+1}\right\rangle) + D_\xi(z, z^k)$$
$$\geq \alpha_{k+1}(g(z^{k+1}) + f(x^{k+1}) + \left\langle df(x^{k+1}), z^{k+1} - x^{k+1}\right\rangle) + D_\xi(z^{k+1}, z^k) + D_\xi(z, z^{k+1})$$

Now we use that $\alpha_{k+1} = \beta_{k+1} - \beta_k$ and $\alpha_{k+1} z^{k+1} = \beta_{k+1} y^{k+1} - \beta_k y^k$ to write:

$$\alpha_{k+1}(f(x^{k+1}) + \left\langle df(x^{k+1}), z^{k+1} - x^{k+1}\right\rangle)$$
$$= \beta_{k+1}(f(x^{k+1}) + \left\langle df(x^{k+1}), y^{k+1} - x^{k+1}\right\rangle)) - \beta_k(f(x^{k+1}) + \left\langle df(x^{k+1}), y^k - x^{k+1}\right\rangle))$$
$$\geq \beta_{k+1}(f(y^{k+1}) - D_f(y^{k+1}, x^{k+1})) - \beta_k f(y^k).$$

Also: $\beta_{k+1} g(y^{k+1}) \leq \alpha_{k+1} g(z^{k+1}) + \beta_k g(y^k)$ by convexity.

Accelerated Mirror descent

Hence combining these inequalities we have:

$$\alpha_{k+1}(g(z^{k+1}) + f(x^{k+1}) + \left\langle df(x^{k+1}), z^{k+1} - x^{k+1} \right\rangle) \geq \beta_{k+1}(F(y^{k+1}) - D_f(y^{k+1}, x^{k+1})) - \beta_k F(y^k),$$

and

$$(\beta_{k+1} - \beta_k)F(z) + D_\xi(z, z^k) \geq \beta_{k+1}(F(y^{k+1}) - D_f(y^{k+1}, x^{k+1})) - \beta_k F(y^k) + D_\xi(z^{k+1}, z^k) + D_\xi(z, z^{k+1}),$$

that is:

$$\beta_k(F(y^k) - F(z)) + D_\xi(z, z^k) \geq \beta_{k+1}(F(y^{k+1}) - F(z)) + D_\xi(z, z^{k+1}) - \beta_{k+1} D_f(y^{k+1}, x^{k+1}) + D_\xi(z^{k+1}, z^k).$$

We now show that $D_\xi(z^{k+1}, z^k) \geq \beta_{k+1} D_f(y^{k+1}, x^{k+1})$.

Accelerated Mirror descent

$D_\xi(z^{k+1}, z^k) \geq \beta_{k+1} D_f(y^{k+1}, x^{k+1})$: here we use that f is L-Lipschitz and ξ 1-convex, so that

$$D_\xi(z^{k+1}, z^k) - \beta_{k+1} D_f(y^{k+1}, x^{k+1}) \geq \frac{1}{2}\left(\|z^{k+1} - z^k\|^2 - \beta_{k+1} L \|y^{k+1} - x^{k+1}\|^2\right)$$
$$= \frac{1}{2}\left(\|z^{k+1} - z^k\|^2 - \beta_{k+1} L \frac{\alpha_{k+1}^2}{\beta_{k+1}^2}\|z^{k+1} - z^k\|^2\right) \geq 0$$

by the definition of β_{k+1}.

We deduce:

$$\beta_k \left(F(y^k) - F(z)\right) \leq D_\xi(z, z^0) + \beta_0(F(y^0) - F(z)) = D_\xi(z, z^0).$$

Now, $\alpha_{k+1} = \frac{1+\sqrt{1+4L\beta_k}}{2L}$ and $\beta_{k+1} = \beta_k + \alpha_{k+1}$. By induction we deduce that $\beta_k \geq k^2/(4L)$. Indeed, if true, it implies:

$$\alpha_{k+1} \geq \frac{1+\sqrt{k^2+1}}{2L} \text{ and } \beta_{k+1} \geq \frac{k^2 + 2 + 2\sqrt{k^2+1}}{4L} = \frac{(\sqrt{k^2+1}+1)^2}{4L} \geq \frac{(k+1)^2}{4L}.$$

Accelerated Mirror descent

Remarks

- A "backtracking" technique is available if one does not know L in advance;
- Requires increasing sequence α_k: might become harder and harder to compute as k increases;
- Better rate if g is *relatively* strongly convex (or f, possibly modifying the algorithm). Linear with $\omega \approx 1 - \sqrt{\mu/L}$ if $\mu << L$ (with varying or fixed α, β);
- "Relatively" strongly convex might not be *very* interesting in general. (Main example: "smoothing".)

Nonlinear primal-dual algorithm

One can extend also the primal-dual algorithm to the non-linear case. In fact, it is even simpler. We introduce strongly convex Legendre functions ξ_x, ξ_y for both x and y and assume we want to solve

$$\min_{x \in \operatorname{dom} \xi_x} \sup_{y \in \operatorname{dom} \xi_y} g(x) + \langle y, Kx \rangle - f^*(y).$$

Algorithm: Bregman PDHG

$$x^{k+1} = \arg\min g(x) + \left\langle y^k, Kx \right\rangle + \frac{1}{\tau} D_x(x, x^k),$$

$$y^{k+1} = \arg\min f^*(y) - \left\langle y, K(2x^{k+1} - x^k) \right\rangle + \frac{1}{\sigma} D_y(y, y^k)$$

Nonlinear primal-dual algorithm: descent rule

With the same notation as in the previous lecture:

$$\hat{y} = \arg\min_y f^*(y) - \langle y, K\tilde{x}\rangle + \frac{1}{\sigma} D_y(y, \bar{y}),$$
$$\hat{x} = \arg\min_x g(x) + \langle \tilde{y}, Kx\rangle + \frac{1}{\tau} D_x(x, \bar{x})$$

we can deduce the same descent rule: for all $x \in \operatorname{dom} \xi_x$, $y \in \operatorname{dom} \xi_y$, one has:

$$g(x) + \langle Kx, \tilde{y}\rangle + \frac{1}{\tau} D_x(x, \bar{x}) \geq g(\hat{x}) + \langle K\hat{x}, \tilde{y}\rangle + \frac{1}{\tau} D_x(\hat{x}, \bar{x}) + \frac{1+\tau\mu_g}{\tau} D_x(x, \hat{x})$$
$$f^*(y) - \langle K\tilde{x}, y\rangle + \frac{1}{\sigma} D_y(y, \bar{y}) \geq f^*(\hat{y}) - \langle K\tilde{x}, \hat{y}\rangle + \frac{1}{\sigma} D_y(\hat{y}, \bar{y}) + \frac{1+\sigma\mu_{f^*}}{\sigma} D_y(y, \hat{y}).$$

reproducing the same computation and using the 3-points inequality (here if g is μ_g relatively strongly convex wr ξ_x, and f^* is μ_{f^*} relatively strongly convex wr ξ_y). Then the convergence proofs are identical. For instance, we get:

Nonlinear primal-dual algorithm

Rate for Nonlinear PDHG

We let $Z^N = (X^N, Y^N)^T := \frac{1}{N}\sum_{k=1}^N z^k$. Then for all $x \in \operatorname{dom}\xi_x$ and $y \in \operatorname{dom}\xi_y$:

$$\mathcal{L}(X^N, y) - \mathcal{L}(x, Y^N) \le \frac{1}{N}\left(\frac{1}{\tau}D_x(x, x^0) + \frac{1}{\sigma}D_y(y, y^0) - \left\langle y - y^0, K(x - x^0)\right\rangle\right)$$

provided $\sigma\tau L^2 \le 1$, where $L := \sup_{\|x\|\le 1, \|y\|\le 1}\langle y, Kx\rangle$.

Remark: under this condition, one has $\langle y - y^0, K(x - x^0)\rangle \le D_x(x, x^0)/\tau + D_y(y, y^0)/\sigma$ so that one can also bound the rate by

$$\cdots \le \frac{2}{N}\left(\frac{1}{\tau}D_x(x, x^0) + \frac{1}{\sigma}D_y(y, y^0)\right).$$

(Accelereated) Nonlinear primal-dual algorithm

If in addition g is μ_g relatively strongly convex, then, as in the Euclidean case, one can update y^k with $x^k + \theta_k(x^k - x^{k-1})$ and then x^k with y^{k+1} and we obtain:

Accelerated rate

Choosing $x^{-1} = x^0$, $\sigma_0 \tau_0 L^2 \le 1$ and for $k \ge 0$, $\theta_{k+1} = 1/\sqrt{1 + \mu_g \tau_k}$, $\tau_{k+1} = \tau_k \theta_{k+1}$, $\sigma_{k+1} = \sigma_k / \theta_{k+1}$, one has:

$$T_N(\mathcal{L}(X^N, y) - \mathcal{L}(x, Y^N)) + \frac{\sigma_N}{2\tau_N}\|x^N - x\|^2 \le \frac{\sigma_0}{\tau_0} D_x(x, x^0) + D_y(y, y^0)$$

where $T_N = \sum_{k=0}^{N-1} \sigma_k \approx \mu_g k^2 / L^2$, $Z^N = \frac{1}{T_N} \sum_{k=0}^{N-1} \sigma_k z^{k+1}$ $(z = (x, y))$.

Application of Bregman (primal-dual) descent

Example: Complexity for "optimal transportation" problems.
Problem: optimal assignment:

$$\min \left\{ C\!:\!X \ :\ X\mathbf{1} = \tfrac{1}{N}\mathbf{1}, X^T\mathbf{1} = \tfrac{1}{N}\mathbf{1}, X \geq 0 \right\}$$

where C is an $N \times N$ cost matrix (in general ≥ 0 but this is not important), X is an $N \times N$ matrix with $\sum_{i,j} X_{i,j} = 1$, $C:X := \sum_{i,j} C_{i,j}X_{i,j}$ and $\mathbf{1} = (1, \dots, 1)^T$. Then one can show that this problem is solved by a permutation matrix $X_{i,j} = \delta_{\epsilon(i),j}$ for $\epsilon \in \mathcal{S}(N)$, which minimizes the cost $\sum_j C_{i,\epsilon(i)}$. More general problem: $X\mathbf{1} = \mu$, $X^T\mathbf{1} = \nu$ where μ, ν are discretized probability measures ($\sum_i \mu_i = 1$): convexification of "optimal transportation" problem (then X might not be a permutation anymore).

Optimal assignment

Primal-dual and dual formulation:

$$\min_{X\geq 0} \sup_{f,g\in\mathbb{R}^N} C : X + f\cdot(\mu - X\mathbf{1}) + g\cdot(\nu - X^T\mathbf{1})$$
$$= \max_{f,g} f\cdot\mu + g\cdot\nu + \min_{X\geq 0} X : (C - f\otimes 1 - 1\otimes g) = \max_{f,g: f_i+g_j\leq C_{i,j}} f\cdot\mu + g\cdot\nu.$$

Then, one can show that there is a solution (X^*, f^*, g^*) with:

$$X_{i,j} > 0 \Rightarrow f_i + g_j = C_{i,j}$$
$$f_i + g_j < C_{i,j} \Rightarrow X_{i,j} = 0.$$

In particular:

- (f, g) solution $\Rightarrow (f + c, g - c)$ solution for any constant c;
- One can find a solution with $|f_i|, |g_j| \leq |C|_\infty/2$ ($|C|_\infty = \max_{i,j} C_{i,j}$).

Optimal assignment

Primal-dual algorithm, for $\lambda = |C|_\infty/2$:

$$\min_{X\geq 0} \sup_{|f|,|g|\leq\lambda} C : X - X : (f\otimes 1 - 1\otimes g) + f\cdot\mu + g\cdot\nu :$$

We pick X^0, f^0, g^0 and let for $k\geq 0$:

$$(f^{k+1}, g^{k+1}) = \arg\min_{|f|,|g|\leq\lambda/2} \frac{1}{\tau}\Big(D_f(f,f^k) + D_f(g,g^k)\Big) - f\cdot\mu - g\cdot\nu - X^k : (f\otimes 1 - 1\otimes g);$$

$$(\bar f^{k+1}, \bar g^{k+1}) = 2(f^{k+1}, g^{k+1}) - (f^k, g^k)$$

$$X^{k+1} = \arg\min_{X\geq 0} \frac{1}{\sigma} D_X(X,X^k) + X : (C - \bar f^{k+1}\otimes 1 - 1\otimes \bar g^{k+1}).$$

(the minimizations wr f and wr g are uncoupled).

Optimal assignment

One obtains a rate of the form:

$$\mathrm{Gap}^k \leq \frac{2}{k}\left(\frac{1}{\sigma}D_X(X,X^0) + \frac{1}{\tau}D_f(f,f^0) + \frac{1}{\tau}D_g(g,g^0)\right).$$

with $\sigma\tau L^2 \leq 1$. Let us consider two cases:

1. $\xi_f = \xi_g = |\cdot|^2/2$, $\xi_X = |\cdot|^2/2$ (Euclidean case);
2. $\xi_f = \xi_g = |\cdot|^2/2$, $\xi_X = \sum_{i,j} X_{i,j} \log X_{i,j}$ with $\sum_{i,j} X_{i,j} = 1$ (Entropy case), and the norm $\|X\| = \|X\|_1 = \sum_{i,j} |X_{i,j}|$.

Optimal assignment
Euclidean

In the first case:

$$L = \sup \left\{ \sum_{i,j} X_{i,j}(f_i + g_j) : \sum_{i,j} X_{i,j}^2 \le 1, \sum_i f_i^2 + g_i^2 \le 1 \right\} = \sup \sqrt{\sum_{i,j} f_i^2 + g_j^2} = \sqrt{N}$$

so one needs $\tau\sigma \le 1/N$. Then, one has (assuming $X^0 = \frac{1}{N^2}\mathbf{1} \otimes \mathbf{1}$ or 0)

$$\sup_{X \ge 0, \sum_{i,j} X_{i,j} = 1} \frac{1}{2}|X - X^0|^2 \le \frac{1}{2}, \quad \sup_{|f|,|g| \le \lambda} \frac{1}{2}(|f|^2 + |g|^2) \le N\lambda^2$$

hence the rate is less than $(2/k)$ times:

$$\min_{\sigma\tau = 1/N} \frac{1}{2\sigma} + \frac{N\lambda^2}{\tau} = \min_{\sigma > 0} \frac{1}{2\sigma} + N^2\lambda^2\sigma = \sqrt{2}N\lambda$$

and the optimum is for $\sigma = 1/(N\lambda\sqrt{2})$, $\tau = \sqrt{2}\lambda$.

Optimal assignment

Non-linear

In the second case:

$$L = \sup \left\{ \sum_{i,j} X_{i,j}(f_i + g_j) : \sum_{i,j} |X_{i,j}| \le 1, \sum_i f_i^2 + g_i^2 \le 1 \right\} = \sup \max_{i,j} f_i + g_j = \sqrt{2}$$

so one needs $\tau\sigma \le 1/2$. One recalls that (for $\sum_{i,j} X_{i,j} = \sum_{i,j} Y_{i,j} = 1$):

$$D_X(X,Y) = \sum_{i,j} X_{i,j} \log X_{i,j} - Y_{i,j} \log Y_{i,j} - (\log Y_{i,j} + 1)(X_{i,j} - Y_{i,j}) = \sum_{i,j} X_{i,j} \log \frac{X_{i,j}}{Y_{i,j}}$$

so that one has (assuming $X^0 = \frac{1}{N^2}\mathbf{1} \otimes \mathbf{1}$)

$$\sup_{X \ge 0, \sum_{i,j} X_{i,j} = 1} \sum_{i,j} X_{i,j} \log \frac{X_{i,j}}{X^0_{i,j}} \le \log N^2.$$

Hence, the rate is less than $(2/k)$ times:

$$\min_{\sigma\tau = 1/2} \frac{2 \log N}{\sigma} + \frac{N\lambda^2}{\tau} = \min_{\sigma > 0} \frac{2 \log N}{\sigma} + 2N\lambda^2\sigma = \sqrt{N \log N}\lambda$$

and the optimum is for $\sigma = \sqrt{\log N/N}/\lambda$, $\tau = (\lambda/2)\sqrt{N/\log N}$.