Beyond metric settings, gradient descent and flow with c(x,y) cost

Pierre-Cyril Aubin

Researcher CERMICS (Optimisation team), ENPC

in collaboration with

Flavien Léger (INRIA Paris)

Giacomo Enrico Sodini, Ulisse Stefanelli (Uni Vienna)

11/03/2025 GdT LPSM Paris-Cité

My type of questions so far: what are the relations between

- concepts, e.g. kernels Hilbertian or tropical
- objective functions f and geometry c
- optimization algorithms, e.g. mirror and natural gradient descent

My type of questions so far: what are the relations between

- concepts, e.g. kernels Hilbertian or tropical
- objective functions f and geometry c
- optimization algorithms, e.g. mirror and natural gradient descent

For today, essentially:

- many discrete-time descent algorithms look similar, can they be unified to study them together?

  $\hookrightarrow$ yes, through alternating minimization (AM)

- the continuous-time formulation of gradient flows has been extended to metric spaces, can we go beyond d^2?

  $\hookrightarrow$ yes, with general costs, when it's about evolution variational inequalities (EVI)

What are we going to see today?

1. Motivation from discrete-time

2. Alternating minimization and GradDesc with GenCost

3. c-EVI and continuous-time

Motivation 1: extending implicit gradient descent and EVIs

Take a C^1 function $g : \mathbb{R}^d \to \mathbb{R}$, $\tau > 0$ and consider the implicit gradient descent

$$x_{n+1} - x_n = -\tau \nabla g(x_{n+1}). \tag{1}$$

It is trivially an alternating minimization of a $\phi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$

$$g(x) \leq \phi(x, x') := g(x) + \frac{1}{2\tau}\|x - x'\|^2. \tag{2}$$

Motivation 1: extending implicit gradient descent and EVIs

Take a C^1 function $g : \mathbb{R}^d \to \mathbb{R}$, $\tau > 0$ and consider the implicit gradient descent

$$x_{n+1} - x_n = -\tau \nabla g(x_{n+1}). \tag{1}$$

It is trivially an alternating minimization of a $\phi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$

$$g(x) \leq \phi(x, x') := g(x) + \frac{1}{2\tau}\|x - x'\|^2. \tag{2}$$

When $\tau \to 0$, we get the *gradient flow* $x'(t) = -\nabla g(x_t)$, or, if g is convex, we have the equivalent *evolution variational inequality* (EVI)

$$\frac{d}{dt}\left(\frac{\|x_t - x\|^2}{2}\right) \leq g(x) - g(x_t) \quad \forall t \in (0, +\infty),\ x \in \mathbb{R}^d$$

obtained as a limit of the discrete EVI: $\frac{\|x_{n+1}-x\|^2}{2\tau} - \frac{\|x_n - x\|^2}{2\tau} + \frac{\|x_{n+1}-x_n\|^2}{2\tau} \leq g(x) - g(x_{n+1})$.

Motivation 1: extending implicit gradient descent and EVIs

Take a C^1 function $g : \mathbb{R}^d \to \mathbb{R}$, $\tau > 0$ and consider the implicit gradient descent

$$x_{n+1} - x_n = -\tau \nabla g(x_{n+1}). \tag{1}$$

It is trivially an alternating minimization of a $\phi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$

$$g(x) \leq \phi(x, x') := g(x) + \frac{1}{2\tau}\|x - x'\|^2. \tag{2}$$

When $\tau \to 0$, we get the *gradient flow* $x'(t) = -\nabla g(x_t)$, or, if g is convex, we have the equivalent *evolution variational inequality* (EVI)

$$\frac{d}{dt}\left(\frac{\|x_t - x\|^2}{2}\right) \leq g(x) - g(x_t) \quad \forall t \in (0, +\infty),\, x \in \mathbb{R}^d$$

obtained as a limit of the discrete EVI: $\frac{\|x_{n+1}-x\|^2}{2\tau} - \frac{\|x_n - x\|^2}{2\tau} + \frac{\|x_{n+1}-x_n\|^2}{2\tau} \leq g(x) - g(x_{n+1})$.

How to generalize this setting when $\|x-x'\|^2/2\tau$ is "replaced" by $c_\tau(x, y)$?

Motivation 2: extending explicit gradient descent

Take a C^2 function $f : \mathbb{R}^d \to \mathbb{R}$, $\tau > 0$ and consider the explicit gradient descent

$$x_{n+1} - x_n = -\tau \nabla f(x_n). \tag{3}$$

To have $\|\nabla f(x_n)\| \xrightarrow{n \to \infty} 0$, $1/\tau$-smoothness ($\nabla^2 f \leq 1/\tau \, \mathrm{Id}$) suffices, as a "descent lemma"

$$f(x') \leq \phi(x, x') := f(x) + \langle \nabla f(x), x' - x \rangle + \frac{1}{2\tau}\|x - x'\|^2. \tag{4}$$

Gradient descent is just minimization of the upper bound!

Motivation 2: extending explicit gradient descent

Take a C^2 function $f : \mathbb{R}^d \to \mathbb{R}$, $\tau > 0$ and consider the explicit gradient descent

$$x_{n+1} - x_n = -\tau \nabla f(x_n). \tag{3}$$

To have $\|\nabla f(x_n)\| \overset{n\to\infty}{\longrightarrow} 0$, $1/\tau$-smoothness ($\nabla^2 f \leq 1/\tau\,\mathrm{Id}$) suffices, as a "descent lemma"

$$f(x') \leq \phi(x, x') := f(x) + \langle \nabla f(x), x' - x\rangle + \frac{1}{2\tau}\|x - x'\|^2. \tag{4}$$

Gradient descent is just minimization of the upper bound!

To obtain (sub)linear convergence of $f(x_n)$, we use (strong) convexity, i.e. for a $\lambda \geq 0$

$$f(x) + \langle \nabla f(x), x' - x\rangle + \frac{\lambda}{2}\|x - x'\|^2 \leq f(x'). \tag{5}$$

There are three objects: i) an algorithm; ii) a regularizer; iii) a class of functions

Motivation 2: extending explicit gradient descent

Take a C^2 function $f : \mathbb{R}^d \to \mathbb{R}$, $\tau > 0$ and consider the explicit gradient descent

$$x_{n+1} - x_n = -\tau \nabla f(x_n). \tag{3}$$

To have $\|\nabla f(x_n)\| \overset{n \to \infty}{\longrightarrow} 0$, ${}^1\!/\!_\tau$-smoothness ($\nabla^2 f \leq {}^1\!/\!_\tau \,\mathrm{Id}$) suffices, as a "descent lemma"

$$f(x') \leq \phi(x, x') := f(x) + \langle \nabla f(x), x' - x \rangle + \frac{1}{2\tau}\|x - x'\|^2. \tag{4}$$

Gradient descent is just minimization of the upper bound!

To obtain (sub)linear convergence of $f(x_n)$, we use (strong) convexity, i.e. for a $\lambda \geq 0$

$$f(x) + \langle \nabla f(x), x' - x \rangle + \frac{\lambda}{2}\|x - x'\|^2 \leq f(x'). \tag{5}$$

There are three objects: i) an algorithm; ii) a regularizer; iii) a class of functions
How are they related? Can we get an EVI for the explicit case too?

Systematic majorization–minimization with a cost

Let $f, g \colon X \to \mathbb{R}$ where X is any set. Choose another set Y and a function $c : X \times Y \to \mathbb{R} \cup \{+\infty\}$. Define the upperbound

$$f(x) + g(x) \leq \phi(x, y) \coloneqq g(x) + c(x, y) + \underbrace{\sup_{x' \in X} [f(x') - c(x', y)]}_{=: f^c(y)} \tag{6}$$

Do alternating minimization (AM) of the surrogate

$$y_{n+1} = \operatorname*{argmin}_{y \in Y} c(x_n, y) + f^c(y) + g(x_n), \tag{7}$$

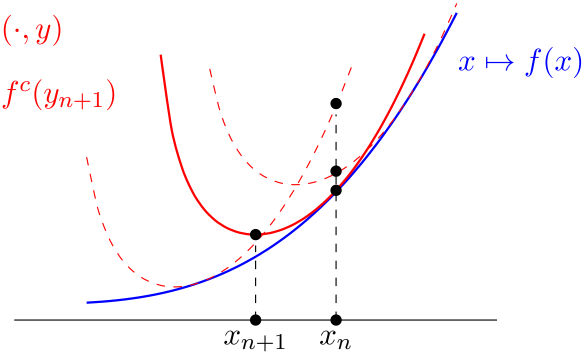
$$x_{n+1} = \operatorname*{argmin}_{x \in X} c(x, y_{n+1}) + f^c(y_{n+1}) + g(x). \tag{8}$$

No topological requirements! Just existence and uniqueness of iterates (always assumed in this talk)

Visual sketch of alternating minimization for $g = 0$

among the upperbounds $\phi(\cdot, y)$

$x \mapsto \phi(x, y_{n+1}) = c(x, y_{n+1}) + f^c(y_{n+1})$



If $f(x) = \inf_y \phi(x, y)$, then $\inf_x f(x) = \inf_{x,y} \phi(x, y)$

Systematic majorization–minimization with a cost

Let $f, g\colon X \to \mathbb{R}$ where X is any set. Choose another set Y and a function $c(x,y)$. Define the upperbound

$$f(x) + g(x) \le \phi(x,y) \coloneqq g(x) + c(x,y) + f^c(y) \coloneqq g(x) + c(x,y) + \sup_{x' \in X} f(x') - c(x',y) \tag{9}$$

Do alternating minimization (AM) of the surrogate

$$y_{n+1} = \operatorname*{argmin}_{y \in Y} c(x_n, y) + f^c(y) + g(x_n), \tag{10}$$

$$x_{n+1} = \operatorname*{argmin}_{x \in X} c(x, y_{n+1}) + f^c(y_{n+1}) + g(x). \tag{11}$$

Systematic majorization–minimization with a cost

Let $f, g\colon X \to \mathbb{R}$ where X is any set. Choose another set Y and a function $c(x,y)$. Define the upperbound

$$f(x) + g(x) \le \phi(x,y) \coloneqq g(x) + c(x,y) + f^c(y) \coloneqq g(x) + c(x,y) + \sup_{x' \in X} f(x') - c(x',y) \tag{9}$$

Do alternating minimization (AM) of the surrogate

$$y_{n+1} = \operatorname*{argmin}_{y \in Y} c(x_n, y) + f^c(y) + g(x_n), \tag{10}$$

$$x_{n+1} = \operatorname*{argmin}_{x \in X} c(x, y_{n+1}) + f^c(y_{n+1}) + g(x). \tag{11}$$

If the setting allows to differentiate and $f(x) = f^{cc}(x) = \inf_y c(x,y) + f^c(y)$ (c-concavity) then we can write (applying the envelope theorem $\nabla f(x) = \nabla_1 \phi(x, \bar{y}(x))$)

$$-\nabla_x c(x_n, y_{n+1}) = -\nabla f(x_n), \tag{12}$$

$$\nabla_x c(x_{n+1}, y_{n+1}) = -\nabla g(x_{n+1}). \tag{13}$$

For a quadratic c, we recover forward–backward gradient descent!

Gradient descent with a general cost - Examples $g = 0$

$$-\nabla_x c(x_n, y_{n+1}) = -\nabla f(x_n), \qquad (y\text{-update})$$
$$\nabla_x c(x_{n+1}, y_{n+1}) = 0. \qquad (x\text{-update})$$

In the following: $Y = X$, and c is minimal on the diagonal $\{x = y\}$, so $x_{n+1} = y_{n+1}$

Gradient descent	$\frac{L}{2}\|x - y\|^2$	$x_{n+1} - x_n = -\frac{1}{L}\nabla f(x_n)$
Mirror descent[1]	$u(x\|y)$	$\nabla u(x_{n+1}) - \nabla u(x_n) = -\nabla f(x_n)$

[1] Bregman divergence of $u : X \to \mathbb{R}$ convex and differentiable is $u(x|y) := u(x) - u(y) - \langle \nabla u(y), x - y \rangle$. E.g. the Kullback–Leibler divergence $\mathrm{KL}(x, y) = \Sigma_i x_i \ln(x_i/y_i)$ for the entropy $u(x) = \Sigma_i x_i \ln(x_i)$ over the simplex.

Gradient descent with a general cost - Examples $g = 0$

$$-\nabla_x c(x_n, y_{n+1}) = -\nabla f(x_n), \qquad (y\text{-update})$$

$$\nabla_x c(x_{n+1}, y_{n+1}) = 0. \qquad (x\text{-update})$$

In the following: $Y = X$, and c is minimal on the diagonal $\{x = y\}$, so $x_{n+1} = y_{n+1}$

Gradient descent	$\frac{L}{2}\|x-y\|^2$	$x_{n+1} - x_n = -\frac{1}{L}\nabla f(x_n)$
Mirror descent[1]	$u(x\|y)$	$\nabla u(x_{n+1}) - \nabla u(x_n) = -\nabla f(x_n)$
Natural gradient descent	$u(y\|x)$	$x_{n+1} - x_n = -(\nabla^2 u(x_n))^{-1}\nabla f(x_n)$
Pre-conditionned gradient descent	$\ell(x-y)$	$x_{n+1} - x_n = -\nabla \ell^*(\nabla f(x_n))$
Riemannian gradient descent	$\frac{L}{2} d_M^2(x,y)$	$x_{n+1} = \exp_{x_n}(-\frac{1}{L}\nabla f(x_n))$

[1] Bregman divergence of $u : X \to \mathbb{R}$ convex and differentiable is $u(x|y) := u(x) - u(y) - \langle \nabla u(y), x - y\rangle$. E.g. the Kullback–Leibler divergence $\mathrm{KL}(x,y) = \Sigma_i x_i \ln(x_i/y_i)$ for the entropy $u(x) = \Sigma_i x_i \ln(x_i)$ over the simplex.

Gradient descent with a general cost - Examples $g = 0$

$$-\nabla_x c(x_n, y_{n+1}) = -\nabla f(x_n), \qquad (y\text{-update})$$
$$\nabla_x c(x_{n+1}, y_{n+1}) = 0. \qquad (x\text{-update})$$

In the following: $Y = X$, and c is minimal on the diagonal $\{x = y\}$, so $x_{n+1} = y_{n+1}$

Gradient descent	$\frac{L}{2}\|x-y\|^2$	$x_{n+1} - x_n = -\frac{1}{L}\nabla f(x_n)$
Mirror descent[1]	$u(x\|y)$	$\nabla u(x_{n+1}) - \nabla u(x_n) = -\nabla f(x_n)$
Natural gradient descent	$u(y\|x)$	$x_{n+1} - x_n = -(\nabla^2 u(x_n))^{-1}\nabla f(x_n)$
Pre-conditionned gradient descent	$\ell(x-y)$	$x_{n+1} - x_n = -\nabla \ell^*(\nabla f(x_n))$
Riemannian gradient descent	$\frac{L}{2}d_M^2(x,y)$	$x_{n+1} = \exp_{x_n}(-\frac{1}{L}\nabla f(x_n))$

We now provide assumptions on f and c to obtain a (sub)linear convergence rate.

[1] Bregman divergence of $u : X \to \mathbb{R}$ convex and differentiable is $u(x|y) \coloneqq u(x) - u(y) - \langle \nabla u(y), x - y\rangle$. E.g. the Kullback–Leibler divergence $\mathrm{KL}(x,y) = \Sigma_i x_i \ln(x_i/y_i)$ for the entropy $u(x) = \Sigma_i x_i \ln(x_i)$ over the simplex.

Convergence rates for the explicit case $g = 0$

Consider the sequence of AM iterates, starting from any x_0,

$$y_n \to x_n \to y_{n+1}$$

We say that f is c-cross-convex if f dominates a *cross-difference* (McCann, 1999)

$$f(x) - f(x_n) \geq c(x, y_{n+1}) - c(x, y_n) + c(x_n, y_n) - c(x_n, y_{n+1}) \, \forall \, x, y_n \in X \times Y.$$

e.g. 3-point inequality (c Bregman), discrete EVI (c Riemann), specific Lyapunov function. . .

Convergence rates for the explicit case $g = 0$

Consider the sequence of AM iterates, starting from any x_0,

$$y_n \to x_n \to y_{n+1}$$

We say that f is c-cross-convex if f dominates a *cross-difference* (McCann, 1999)

$$f(x) - f(x_n) \geq c(x, y_{n+1}) - c(x, y_n) + c(x_n, y_n) - c(x_n, y_{n+1}) \, \forall \, x, y_n \in X \times Y.$$

c-concavity ($f(x) = \inf_y c(x, y) + f^c(y)$) implies, since $f^c(y_{n+1}) = f(x_n) - c(x_n, y_{n+1})$,

$$f(x) - f(x_n) \leq c(x, y_{n+1}) - c(x_n, y_{n+1}).$$

Convergence rates for the explicit case $g = 0$

Consider the sequence of AM iterates, starting from any x_0,

$$y_n \to x_n \to y_{n+1}$$

We say that f is c-cross-convex if f dominates a *cross-difference* (McCann, 1999)

$$f(x) - f(x_n) \geq c(x, y_{n+1}) - c(x, y_n) + c(x_n, y_n) - c(x_n, y_{n+1}) \, \forall \, x, y_n \in X \times Y.$$

c-concavity $(f(x) = \inf_y c(x,y) + f^c(y))$ implies, since $f^c(y_{n+1}) = f(x_n) - c(x_n, y_{n+1})$,

$$f(x) - f(x_n) \leq c(x, y_{n+1}) - c(x_n, y_{n+1}).$$

For $c(x,y) = \frac{L}{2}\|x-y\|^2$ and $f \in C^1(\mathbb{R}^d, \mathbb{R})$, we get $x_n = y_n$ and $x_{n+1} - x_n = -\tau \nabla f(x_n)$

$$\langle \nabla f(x_n), x - x_n \rangle \leq f(x) - f(x_n) \leq \frac{L}{2}\|x - x_{n+1}\|^2 - \frac{1}{2L}\|\nabla f(x_n)\|^2 = \langle \nabla f(x_n), x - x_n \rangle + \frac{L}{2}\|x - x_n\|^2$$

Convergence rates for the explicit case $g = 0$

Consider the sequence of AM iterates, starting from any x_0,

$$y_n \to x_n \to y_{n+1}$$

We say that f is c-cross-convex if f dominates a *cross-difference* (McCann, 1999)

$$f(x) - f(x_n) \geq c(x, y_{n+1}) - c(x, y_n) + c(x_n, y_n) - c(x_n, y_{n+1}) \, \forall \, x, y_n \in X \times Y.$$

c-concavity ($f(x) = \inf_y c(x,y) + f^c(y)$) implies, since $f^c(y_{n+1}) = f(x_n) - c(x_n, y_{n+1})$,

$$f(x) - f(x_n) \leq c(x, y_{n+1}) - c(x_n, y_{n+1}).$$

For $c(x,y) = \frac{L}{2}\|x-y\|^2$ and $f \in C^1(\mathbb{R}^d, \mathbb{R})$, we get $x_n = y_n$ and $x_{n+1} - x_n = -\tau \nabla f(x_n)$

$$\langle \nabla f(x_n), x - x_n \rangle \leq f(x) - f(x_n) \leq \frac{L}{2}\|x - x_{n+1}\|^2 - \frac{1}{2L}\|\nabla f(x_n)\|^2 = \langle \nabla f(x_n), x - x_n \rangle + \frac{L}{2}\|x - x_n\|^2$$

Suppose that f is c-concave and c-cross-convex, and $x_* = \operatorname{argmin}_X f$. Then

$$f(x_n) - f(x_*) \leq \frac{c(x_*, y_0) - c(x_0, y_0)}{n}. \tag{14}$$

Linear rates and local characterization of c-concavity and c-cross-convexity also exist.

Alternating minimization (AM)

Let $\phi(x, y)\colon X \times Y \to \mathbb{R}$ where X, Y are any sets. Perform an alternating minimization

$$\begin{aligned} y_{n+1} &= \operatorname*{argmin}_{y \in Y} \phi(x_n, y) \\ x_{n+1} &= \operatorname*{argmin}_{x \in X} \phi(x, y_{n+1}), \end{aligned} \tag{15}$$

Many algorithms are AM: alternating projections, Sinkhorn/IPFP, EM,...

Definition (Five-point property (FP) inspired by [Csiszár and Tusnády, 1984])

For $\lambda \geq 0$, ϕ has the λ-FP if $\forall y_0 \in Y$, $\exists x_0, y_1$ s.t. $\forall x, y$

$$\phi(x, y_1) + (1-\lambda)\phi(x_0, y_0) \leq \phi(x, y) + (1-\lambda)\phi(x, y_0). \tag{λ-FP}$$

Note that (λ-FP) forces that $y_0 \to x_0 \to y_1$ as in (15).

$$\phi(x, y_1) + (1-\lambda)\phi(x_0, y_0) \le \phi(x, y) + (1-\lambda)\phi(x, y_0). \tag{λ-FP}$$

Theorem (Convergence rates for alternating minimization)

Suppose that ϕ has a minimizer. Then:

i) For all $n \ge 0$, $\phi(x_{n+1}, y_{n+1}) \le \phi(x_n, y_{n+1}) \le \phi(x_n, y_n)$.

ii) If ϕ satisfies (λ-FP) for $\lambda = 0$. Then for any $x \in X, y \in Y$ and any $n \ge 1$,

$$\phi(x_n, y_n) \le \phi(x, y) + \frac{\phi(x, y_0) - \phi(x_0, y_0)}{n}, \qquad \text{so } \phi(x_n, y_n) - \phi_* = O(1/n)$$

iii) If ϕ satisfies (λ-FP) for some $\lambda \in (0, 1)$. Then for any $x \in X, y \in Y$ and any $n \ge 1$,

$$\phi(x_n, y_n) \le \phi(x, y) + \frac{\lambda[\phi(x, y_0) - \phi(x_0, y_0)]}{\Lambda^n - 1},$$

where $\Lambda := (1-\lambda)^{-1} > 1$. In particular $\phi(x_n, y_n) - \phi_* = O((1-\lambda)^n)$.

Proof

$$\phi(x, y_{n+1}) + \phi(x_n, y_n) \leq \phi(x, y) + \phi(x, y_n). \tag{0-FP}$$

(i): $\phi(x_{n+1}, y_{n+1}) \leq \phi(x_n, y_{n+1}) \leq \phi(x_n, y_n)$ by definition of the iterates.
(ii): (0-FP) can be written as

$$\phi(x_{n+1}, y_{n+1}) \leq \phi(x, y) + [\phi(x, y_n) - \phi(x_n, y_n)] - [\phi(x, y_{n+1}) - \phi(x_{n+1}, y_{n+1})].$$

The last terms inside the brackets are nonnegative. Sum from 0 to $n-1$ and use (i):

$$n\phi(x_n, y_n) \leq \sum_{k=0}^{n-1} \phi(x_{k+1}, y_{k+1}) \leq n\phi(x, y) + [\phi(x, y_0) - \phi(x_0, y_0)] - [\phi(x, y_n) - \phi(x_n, y_n)],$$

Proof

$$\phi(x, y_{n+1}) + \phi(x_n, y_n) \le \phi(x, y) + \phi(x, y_n). \tag{0-FP}$$

(i): $\phi(x_{n+1}, y_{n+1}) \le \phi(x_n, y_{n+1}) \le \phi(x_n, y_n)$ by definition of the iterates.
(ii): (0-FP) can be written as

$$\phi(x_{n+1}, y_{n+1}) \le \phi(x, y) + [\phi(x, y_n) - \phi(x_n, y_n)] - [\phi(x, y_{n+1}) - \phi(x_{n+1}, y_{n+1})].$$

The last terms inside the brackets are nonnegative. Sum from 0 to $n-1$ and use (i):

$$n\phi(x_n, y_n) \le \sum_{k=0}^{n-1} \phi(x_{k+1}, y_{k+1}) \le n\phi(x, y) + [\phi(x, y_0) - \phi(x_0, y_0)] - [\phi(x, y_n) - \phi(x_n, y_n)],$$

[Csiszár and Tusnády, 1984] had given a similar formula, shown convergence to ϕ_* but ... had not seen the convergence rate!

Theorem (Corollary/Convergence rates for GD with general cost)

i) *Suppose that f is c-concave. Then we have the descent property+stopping criterion*

$$f(x_{n+1}) \leq f(x_n) - [c(x_n, y_{n+1}) - c(x_{n+1}, y_{n+1})] \leq f(x_n),$$

$$\min_{0\leq k\leq n-1} [c(x_k, y_{k+1}) - c(x_{k+1}, y_{k+1})] \leq \frac{f(x_0) - f_*}{n}.$$

ii) *Suppose in addition that f is c-cross-convex. Then for any $x \in X, n \geq 1$,*

$$f(x_n) \leq f(x) + \frac{c(x, y_0) - c(x_0, y_0)}{n}. \tag{16}$$

iii) *Suppose in addition that f is λ-strongly c-cross-convex for some $\lambda \in (0, 1)$. Then for any $x \in X, n \geq 1$, setting $\Lambda := (1 - \lambda)^{-1} > 1$*

$$f(x_n) \leq f(x) + \frac{\lambda\,(c(x, y_0) - c(x_0, y_0))}{\Lambda^n - 1}, \tag{17}$$

First primal-primal proof of $1/n$ rate for Sinkhorn with unbounded ground cost. Now, on to EVIs!

For $\phi(x,y) = g(x) + \frac{c(x,y)}{\tau}$ and $c : X \times X \to \mathbb{R}_+$ with $c(x,x) = 0$, the λ-FP reads

$$\phi(x, y_1) + (1-\lambda)\phi(x_0, y_0) \leq \phi(x,y) + (1-\lambda)\phi(x, y_0), \ \forall x, y, y_0 \tag{λ-FP}$$

$$\frac{c(x, x_{n+1}) - c(x, x_n)}{\tau} + \frac{c(x_{n+1}, x_n)}{\tau} + \lambda \frac{c(x, x_n) - c(x_{n+1}, x_n)}{\tau} \leq (1-\lambda)(g(x) - g(x_{n+1})) \ \forall x, y_n$$

For $\lambda = 0$, (X, d) a metric space and $c(x,y) = \frac{d^2(x,y)}{2}$, we get the discrete EVI of [Ambrosio et al., 2008, Corollary 4.1.3]!

For $\phi(x,y) = g(x) + \frac{c(x,y)}{\tau}$ and $c : X \times X \to \mathbb{R}_+$ with $c(x,x) = 0$, the λ-FP reads

$$\phi(x,y_1) + (1-\lambda)\phi(x_0,y_0) \le \phi(x,y) + (1-\lambda)\phi(x,y_0),\ \forall x,y,y_0 \tag{λ-FP}$$

$$\frac{c(x,x_{n+1}) - c(x,x_n)}{\tau} + \frac{c(x_{n+1},x_n)}{\tau} + \lambda\frac{c(x,x_n) - c(x_{n+1},x_n)}{\tau} \le (1-\lambda)(g(x) - g(x_{n+1}))\ \forall x, y_n$$

For $\lambda = 0$, (X,d) a metric space and $c(x,y) = \frac{d^2(x,y)}{2}$, we get the discrete EVI of [Ambrosio et al., 2008, Corollary 4.1.3]!

In practice, for $\mu + 1 = \frac{1}{\tau(1-\lambda)}$, we start from a notion of c/τ-*cross-concavity* of $-g$, i.e.

$$\frac{c(x,x_{n+1}) - c(x,x_n)}{\tau} + \frac{c(x_{n+1},x_n)}{\tau} + \mu c(x,x_n) \le g(x) - g(x_{n+1})\ \forall x,x_n \in X. \tag{18}$$

For $\phi(x,y) = g(x) + \frac{c(x,y)}{\tau}$ and $c : X \times X \to \mathbb{R}_+$ with $c(x,x) = 0$, the λ-FP reads

$$\phi(x, y_1) + (1-\lambda)\phi(x_0, y_0) \le \phi(x,y) + (1-\lambda)\phi(x, y_0),\ \forall x, y, y_0 \tag{λ-FP}$$

$$\frac{c(x,x_{n+1}) - c(x,x_n)}{\tau} + \frac{c(x_{n+1},x_n)}{\tau} + \lambda\frac{c(x,x_n) - c(x_{n+1},x_n)}{\tau} \le (1-\lambda)(g(x) - g(x_{n+1}))\ \forall x, y_n$$

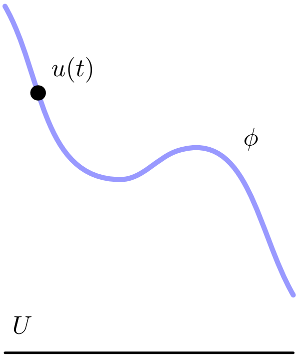
For $\lambda = 0$, (X,d) a metric space and $c(x,y) = \frac{d^2(x,y)}{2}$, we get the discrete EVI of [Ambrosio et al., 2008, Corollary 4.1.3]!
In practice, for $\mu + 1 = \frac{1}{\tau(1-\lambda)}$, we start from a notion of c/τ-*cross-concavity* of $-g$, i.e.

$$\frac{c(x,x_{n+1}) - c(x,x_n)}{\tau} + \frac{c(x_{n+1},x_n)}{\tau} + \mu c(x,x_n) \le g(x) - g(x_{n+1})\ \forall x, x_n \in X. \tag{18}$$

For $\tau \to 0$ and some continuity of g and c, there is a limiting curve satisfying "$\lim_{\tau\downarrow 0} \frac{\nabla_1 c(x_t^\tau, x_t)}{\tau} = -\nabla g(x_t)$" and more precisely a c-EVI:

$$\frac{d}{dt}c(x,x_t) + \mu \cdot c(x,x_t) \le g(x) - g(x_t) \quad \forall t \in (0,+\infty),\ x \in X.$$

Gradient flows



In a Hilbert space X
[Kōmura, Crandall, Pazy, Kato, Brézis, ...]

$$x_t' + \nabla\phi(x_t) = 0$$

- Paradigmatic evolution mode
- Optimization tool

Underlying, there is the squared norm $\|x - y\|^2$

- What if we have just a metric space $d(x, y)$? [Ambrosio, Gigli, Savaré, 2008]
- What if we have just a generic cost $c(x, y)$? [Aubin, Sodini, Stefanelli, 2025?]

Recap of the EVI metric formulation

Unfortunately $x_t' = -\nabla\phi(x_t)$ is not suitable for the metric context (∇ undefined etc).

However, taking inner product with $x_t - x$ where $x \in X$ is arbitrary,

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}\|x_t - x\|^2 = \langle x_t - x, x_t' \rangle = \langle x - x_t, \nabla\phi(x_t)\rangle \leq \phi(x) - \phi(x_t) - \frac{\lambda}{2}\|x_t - x\|^2$$

where we assumed that ϕ is λ-convex, i.e. for all $\bar{x}, x \in X$

$$\langle x - \bar{x}, \nabla\phi(\bar{x})\rangle \leq \phi(x) - \phi(\bar{x}) - \frac{\lambda}{2}\|\bar{x} - x\|^2.$$

Recap of the EVI metric formulation

Unfortunately $x_t' = -\nabla\phi(x_t)$ is not suitable for the metric context (∇ undefined etc).

However, taking inner product with $x_t - x$ where $x \in X$ is arbitrary,

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}\|x_t - x\|^2 = \langle x_t - x, x_t' \rangle = \langle x - x_t, \nabla\phi(x_t)\rangle \leq \phi(x) - \phi(x_t) - \frac{\lambda}{2}\|x_t - x\|^2$$

where we assumed that ϕ is λ-convex, i.e. for all $\bar{x}, x \in X$

$$\langle x - \bar{x}, \nabla\phi(\bar{x})\rangle \leq \phi(x) - \phi(\bar{x}) - \frac{\lambda}{2}\|\bar{x} - x\|^2.$$

Evolution variational inequality:

$x : [0,\infty) \to \operatorname{dom}\phi$ starting from $x^0 \in \operatorname{dom}\phi$ is a EVI solution if

$$\boxed{\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}d^2(x_t, x) + \frac{\lambda}{2}d^2(x_t, x) \leq \phi(x) - \phi(x_t) \quad \text{a.e. } t > 0,\ \forall x} \tag{EVI}$$

Other gradient flow formulations

Defining metric derivative/slope

$$|x_t'| := \lim_{h \to 0^+} \frac{d(x_t, x_{t+h})}{h} \qquad |\nabla\phi|(x) := \max\left(0, \limsup_{y \to x} \frac{\phi(x) - \phi(y)}{d(x,y)}\right)$$

there are two other metric formulations: EDI and EDE, Energy Dissipation (In)Equality

$$\frac{1}{2}\int_s^t |x_t'|^2 \, \mathrm{d}r + \frac{1}{2}\int_s^t |\nabla\phi|^2(x) \, \mathrm{d}r \le \phi(x_s) - \phi(x_t) \tag{EDI}$$

$$\frac{1}{2}\int_s^t |x_t'|^2 \, \mathrm{d}r + \frac{1}{2}\int_s^t |\nabla\phi|^2(x) \, \mathrm{d}r = \phi(x_s) - \phi(x_t) \tag{EDE}$$

These correspond to the energy identity $\frac{\mathrm{d}}{\mathrm{d}t}\phi(x_{t_+}) = -\frac{1}{2}|x_t'|^2 - \frac{1}{2}|\nabla\phi|^2(x) = -|x_t'|^2$
But only the EVI formulation ensures uniqueness and contractivity:

$$\boxed{\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}d^2(x_t, x) + \frac{\lambda}{2}d^2(x_t, x) \le \phi(x) - \phi(x_t) \quad \text{a.e. } t > 0, \ \forall x} \tag{EVI}$$

Glimpse of metric setting literature

EVI in metric setting have been considered in

- Smooth and complete Riemannian manifolds
- Nonpositively curved (NPC) spaces [Mayer, Jost]
- Positively curved (PC) in the Alexandrov sense [Ohta, Savaré, Gigli, Kuwada]
- Wasserstein-Kantorovich-Rubinstein space $(\mathcal{P}_2(X), d_{W2})$ [Ambrosio, Gigli, Savaré, Ohta]
- $RCD(K, \infty)$ spaces [Ambrosio, Gigli, Mondino, Savaré, Erbar, Sturm, Kuwada]

and also extended/adapted to cover

- Reaction-diffusion equations and systems [Kondratyev, Monsaingeon, Vorotnikov, Liero, Mielke, Savaré]
- Viscoelasticity [Mielke, Ortner, Sengül, Friedrich, Kružík]
- Markov chains [Maas, Mielke]
- Jump processes [Erbar, Tse, Rossi, Savaré, Peletier]

General-cost setting

Aim: replace d with a general cost $c : X \times X \to [0, \infty)$

Asymmetric distances have already been considered

[Rossi, Mielke, Savaré, 2008], [Chenchiah, Rieger, Zimmer, 2009]
[Ohta, Zhao, 2024]

For today's presentation I keep:

- symmetry: $c(x, y) = c(y, x)$
- nondegeneracy: $c(x, y) = 0 \;\Leftrightarrow\; x = y$

but I drop the triangle inequality and the continuity of c

(some of our results hold for asymmetric and/or degenerate costs, as well)

General-cost setting: First examples

- Consistency

$$\text{Hilbert: } c(x,y) = \frac{1}{2}\|x-y\|^2, \quad \text{Metric: } c(x,y) = \frac{1}{2}d^2(x,y)$$

- Doubly nonlinear flows

$$c(x,y) = \psi(x-y)$$

  - Continuous problem: $\partial\psi(x') + \partial\phi(x) \ni 0$

- Mirror descent

$$c(x,y) = \psi(x) - \psi(y) - \mathrm{d}\psi(y)(x-y)$$

  - Discrete scheme: $\frac{1}{\tau}(\partial\psi(x_i) - \partial\psi(x_{i-1})) + \partial\phi(x_i) \ni 0$
  - Continuous problem: $(\partial\psi(x))' + \partial\phi(x) \ni 0$

General-cost setting: Examples of interest

- Kullback-Leibler divergence in $\mathcal{P}(X) \times \mathcal{P}(X)$

$$\mathrm{KL}(\mu, \nu) = \begin{cases} \int_X \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(z)\right) \mathrm{d}\mu(z) & \mu \ll \nu \\ \infty & \text{else} \end{cases}$$

- Sinkhorn divergence
  Entropic OT dissimilarity:

$$\mathrm{OT}_\varepsilon(\mu, \nu) = \min_{\pi \in \Pi(\mu,\nu)} \int_{X \times X} c(x, y)\, \mathrm{d}\pi(x, y) + \varepsilon \mathrm{KL}(\pi, \mu \otimes \nu)$$

  Sinkhorn divergence:

$$\mathrm{S}_\varepsilon(\mu, \nu) = \mathrm{OT}_\varepsilon(\mu, \nu) - \frac{1}{2}\mathrm{OT}_\varepsilon(\mu, \mu) - \frac{1}{2}\mathrm{OT}_\varepsilon(\nu, \nu)$$

General-cost setting: Assumptions

- Cost: $c(x,y) = c(y,x) \geq 0$ and $c(x,y) = 0 \;\Leftrightarrow\; x = y$
- Completeness: c-Cauchy sequences are c-convergent, i.e.,

$$c(x_n, x_m) \to 0 \Rightarrow \;\exists \bar{x} \in X \;\, c(x_n, \bar{x}) \to 0$$

- Coercivity: $\forall \tau \in (0,1)$, $\forall y \in X$ the map $x \mapsto c(x,y)/\tau + \phi(x)$ is coercive
- Lower semicontinuity:

$$[c(x_n, x) \to 0] \Rightarrow \;\left[\phi(x) \leq \liminf_n \phi(x_n) \;\text{ and }\; c(x,y) \leq \liminf_n c(x_n, y)\right]$$

- c-cross-convexity: $\forall \tau \in (0,1)$, $\forall x_0,\, u \in \mathrm{dom}(\phi)$, $x_1 \in \arg\min c(\cdot, x_0)/\tau + \phi(\cdot)$

$$\phi(x_1) - \phi(x) \leq \frac{1}{\tau}\left(c(x, x_0) - c(x_1, x_0) - c(x, x_1)\right) - \frac{\lambda}{\tau} c(x, x_1)$$

- Initial value: $x^0 \in \mathrm{dom}(\phi)$

General-cost setting

- A sufficient condition for c-cross-convexity is, for all x, x_0, x_1, the existence of $\gamma : [0,1] \to X$ such that

$$\phi(\gamma(t)) \leq t\phi(x) + (1-t)\phi(x_1) - \lambda t c(x, x_1) + o(t)$$

$$c(\gamma(t), x_0) \leq t c(x, x_0) + (1-t) c(x_1, x_0) - t c(x, x_1) + o(t)$$

- The assumptions are consistent with the metric setting in NPC/NNCC spaces, in particular with Hilbert spaces.
  Think of parallelogram: $\|tx + (1-t)x_1 - x_0\|^2 = t\|x - x_0\|^2 + (1-t)\|x_1 - x_0\|^2 - t(1-t)\|x - x_1\|^2$
- Minimizing Movements, as implicit Euler

$$x_i^\tau \in \arg\min_u \left(\frac{1}{\tau} c(x, x_{i-1}^\tau) + \phi(x) \right)$$

(theory for explicit Euler is also possible)

EVI solution: equivalent formulations

- Differential form:

$$\frac{\mathrm{d}^+}{\mathrm{d}t} c(x_t, x) + \lambda c(x_t, x) \le \phi(x) - \phi(x_t)$$

- Integrated form:

$$c(x_t, x) - c(x_s, x) + \lambda \int_s^t c(x_r, x)\,\mathrm{d}r \le (t-s)\phi(x) - \int_s^t \phi(x_r)\,\mathrm{d}r$$

- Exponential form:

$$\mathrm{e}^{\lambda(t-s)} c(x_t, x) - c(x_s, x) \le \frac{\mathrm{e}^{\lambda(t-s)} - 1}{\lambda}(\phi(x) - \phi(x_t))$$

EVI solution: properties

i) Existence: based on compatibility or c-cross-convexity

ii) Regularizing property and Energy identity: the limits below exist and we have

$$|x'_{t_+}|_c^2 := \lim_{h\to 0+} \frac{2c(x_{t+h}, x_t)}{h^2} \quad , \quad \frac{\mathrm{d}}{\mathrm{d}t}\phi(x_{t_+}) = -|x'_{t_+}|_c^2 \quad \forall t > 0$$

iii) λ-Contractivity (and uniqueness):

$$c(x_t, \tilde{x}_t) \leq \mathrm{e}^{-2\lambda(t-s)} c(x_s, \tilde{x}_s)$$

$\hookrightarrow$ ii) and iii) are a consequence of the symmetry of c ! They do not hold in general.

iv) Large-time behavior:

if $\lambda > 0$ and x_* is the (unique) minimum point of ϕ

$$\frac{\lambda}{2} c(x_t, x_*) \leq \phi(x_t) - \phi(x_*) \leq \lambda \mathrm{e}^{-\lambda t} c(x^0, x_*)$$

v) Stability w.r.t. initial conditions:

$$x_n^0 \to x^0 \quad \Rightarrow \quad x_n(t) \to x_t \quad \forall t > 0$$

Conclusion

- Presented a setting for gradient descent/flow with general costs, consistent with previous metric theory
- EVI solutions introduced & properties discussed, λ-contractivity checked
- Existence for GMM and EVI
- Questions: new PDEs? Novel schemes? Interesting c and ϕ?

Conclusion

- Presented a setting for gradient descent/flow with general costs, consistent with previous metric theory
- EVI solutions introduced & properties discussed, λ-contractivity checked
- Existence for GMM and EVI
- Questions: new PDEs? Novel schemes? Interesting c and ϕ?
- Project originated from article with Flavien Léger (INRIA Paris), more on

https://pcaubin.github.io/

Conclusion

- Presented a setting for gradient descent/flow with general costs, consistent with previous metric theory
- EVI solutions introduced & properties discussed, λ-contractivity checked
- Existence for GMM and EVI
- Questions: new PDEs? Novel schemes? Interesting c and ϕ?
- Project originated from article with Flavien Léger (INRIA Paris), more on

https://pcaubin.github.io/

Thank you for your attention!

arXiv: *Gradient descent with general cost* with *Flavien Léger*

References I

Ambrosio, L., Gigli, N., and Savaré, G. (2008).
Gradient flows in metric spaces and in the space of probability measures.
Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition.

Aubin-Frankowski, P.-C., Korba, A., and Léger, F. (2022).
Mirror descent with relative smoothness in measure spaces, with application to Sinkhorn and EM.
In *Advances in Neural Information Processing Systems (NeurIPS).*
(`https://arxiv.org/abs/2206.08873`).

Bauschke, H. H. and Combettes, P. L. (2011).
Convex analysis and monotone operator theory in Hilbert spaces.
CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, New York.

Csiszár, I. and Tusnády, G. (1984).
Information Geometry and Alternating Minimization Procedures.
In *Statistics and Decisions*, pages 205–237. Oldenburg Verlag, Munich.

References II

Kunstner, F., Kumar, R., and Schmidt, M. W. (2021).
Homeomorphic-invariance of EM: Non-asymptotic convergence in KL divergence for exponential families via mirror descent.
In *AISTATS*.

Neal, R. M. and Hinton, G. E. (1998).
A view of the EM algorithm that justifies incremental, sparse, and other variants.
In *Learning in Graphical Models*, pages 355–368. Springer Netherlands.

Villani, C. (2009).
Optimal transport, volume 338 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*.
Springer-Verlag, Berlin.
Old and new.

c-concavity

Definition (c-concavity)

We say that a function $f\colon X \to \mathbb{R}$ is c-concave if there exists a function $h\colon Y \to \mathbb{R}$ such that

$$f(x) = \inf_{y \in Y} c(x, y) + h(y), \tag{19}$$

for all $x \in X$. If f is c-concave, then we can take $h(y) = f^c(y) = \sup_{x' \in X} f(x') - c(x', y)$.

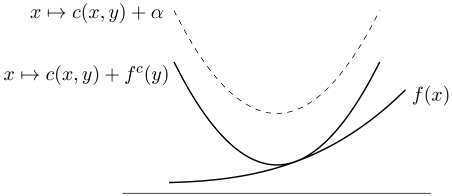


Figure: The c-transform of f. For a fixed $y \in Y$, the dashed line represents a function $x \mapsto c(x, y) + \alpha$ majorizing f. The smallest of such functions is $x \mapsto c(x, y) + f^c(y)$, here represented in solid line.

c-cross-convexity

We want $f(x) - f(x_n) \geq c(x, y_{n+1}) - c(x, y_n) + c(x_n, y_n) - c(x_n, y_{n+1})$ with $-\nabla_x c(x_n, y_{n+1}) = -\nabla f(x_n)$ and $\nabla_x c(x_n, y_n) = 0$.

Recall the *cross-difference* of c defined by

$$\delta_c(x', y'; x, y) := c(x, y') + c(x', y) - c(x, y) - c(x', y').$$

Definition (cross-convexity)

Take f and c C^1. We say that f is c-cross-convex if for all $x, \bar{x} \in X$ and any $\bar{y}, \hat{y} \in Y$ verifying $\nabla_x c(\bar{x}, \bar{y}) = 0$ and $-\nabla_x c(\bar{x}, \hat{y}) = -\nabla f(\bar{x})$ we have

$$f(x) \geq f(\bar{x}) + \delta_c(x, \bar{y}; \bar{x}, \hat{y}). \tag{20}$$

In addition let $\lambda > 0$. We say that f is λ-strongly c-cross-convex if we have

$$f(x) \geq f(\bar{x}) + \delta_c(x, \bar{y}; \bar{x}, \hat{y}) + \lambda(c(x, \bar{y}) - c(\bar{x}, \bar{y})). \tag{21}$$

Local criteria

If $X, Y \subset \mathbb{R}^d$, then we have a local criterion:

Theorem (Local criterion for c-concavity [Villani, 2009, Theorem 12.46])

Suppose that $c \in C^4(X \times Y)$ has nonnegative cross-curvature, $\nabla^2_{xy} c(x, y)$ is everywhere invertible, X and Y have c-segments. Let f be C^2. Suppose that for all $\bar{x} \in X$, there exists $\hat{y} \in Y$ satisfying $-\nabla_x c(\bar{x}, \hat{y}) = -\nabla f(\bar{x})$ and such that

$$\nabla^2 f(\bar{x}) \leq \nabla^2_{xx} c(\bar{x}, \hat{y}).$$

Then f is c-concave. (Converse is also true)

If f is c-cross-convex then, whenever $\nabla_x c(\bar{x}, \bar{y}) = 0$ and $-\nabla_x c(\bar{x}, \hat{y}) = -\nabla f(\bar{x})$, we have

$$\nabla^2 f(\bar{x}) \geq \nabla^2_{xx} c(\bar{x}, \hat{y}) - \nabla^2_{xx} c(\bar{x}, \bar{y}). \tag{22}$$

(Converse is maybe true, a semi-local condition with c-segments does exist though)

POCS (Projection Onto Convex Sets) [Bauschke and Combettes, 2011]

Context: $(H, \|\cdot\|)$ Hilbert space, B, C be two closed convex subsets of H.
Objective: Find $x \in B \cap C$, based on initialization $x_0 \in H$
The POCS algorithm searches for $B \cap C$ by successive projections. Given $x_n \in B$,

$$\begin{aligned} y_{n+1} &= \operatorname*{argmin}_{y \in C} \|x_n - y\|, \\ x_{n+1} &= \operatorname*{argmin}_{x \in B} \|x - y_{n+1}\|. \end{aligned} \tag{23}$$

POCS (Projection Onto Convex Sets) [Bauschke and Combettes, 2011]

Context: $(H, \|\cdot\|)$ Hilbert space, B, C be two closed convex subsets of H.
Objective: Find $x \in B \cap C$, based on initialization $x_0 \in H$
The POCS algorithm searches for $B \cap C$ by successive projections. Given $x_n \in B$,

$$\begin{aligned} y_{n+1} &= \operatorname*{argmin}_{y \in C} \|x_n - y\|, \\ x_{n+1} &= \operatorname*{argmin}_{x \in B} \|x - y_{n+1}\|. \end{aligned} \tag{23}$$

There are at least two ways to write POCS as an alternating minimization method:

i) Take $X = Y = H$, with $c(x, y) = \frac{1}{2}\|x - y\|^2$ and $g = \iota_B$ and $h = \iota_C$, set $\phi(x, y) = c(x, y) + g(x) + h(y)$.

ii) Take $X = B$, $Y = C$ and $\phi(x, y) = \frac{1}{2}\|x - y\|^2$.

In both cases, we can do the analysis to get rates. Same results when $\|x - y\|$ is replaced by $u(x|y)$ (Bregman projections).

Expectation–Maximization (EM)

Context: X : observation space, Z : latent space, Θ : set of parameters, defining our our *statistical models* $\{p_\theta \in \mathcal{P}(\mathsf{X} \times \mathsf{Z}) : \theta \in \Theta\}$.
Objective: Having observed $\mu \in \mathcal{P}(\mathsf{X})$, find $\theta \in \Theta$ maximizing the *likelihood*,

$$\min_{\theta \in \Theta} F(\theta) = \mathrm{KL}(\mu | p_{\mathsf{X}} p_\theta), \tag{24}$$

Use the *data processing inequality*: $F(\theta) = \mathrm{KL}(\mu | p_{\mathsf{X}} p_\theta) \leq \mathrm{KL}(\pi | p_\theta) =: \Phi(\theta, \pi)$.Equality holds for $\pi = \frac{\mu(dx)}{p_{\mathsf{X}} p_\theta(dx)} p_\theta(dx, dz)$. The EM algorithm is [Neal and Hinton, 1998]:

$$\pi_{n+1} = \operatorname*{argmin}_{\pi \in \Pi(\mu, *)} \mathrm{KL}(\pi | p_{\theta_n}), \tag{E-step}$$

$$\theta_{n+1} = \operatorname*{argmin}_{\theta \in \Theta} \mathrm{KL}(\pi_{n+1} | p_\theta). \tag{M-step}$$

It can be written as either mirror descent (convex if $p_\theta = K \otimes \theta$ [Aubin-Frankowski et al., 2022]) or a projected natural gradient descent (convex if p_θ is an exponential family [Kunstner et al., 2021])

Sinkhorn algorithm/Entropic optimal transport

Let (X, μ) and (Y, ν) be two probability spaces and take the set of couplings over $\mathsf{X} \times \mathsf{Y}$ (i.e. joint laws) having marginal μ (resp. ν)

$$C = \Pi(\mu, *), \quad D = \Pi(*, \nu), \quad \Pi(\mu, \nu) = \Pi(\mu, *) \cap \Pi(*, \nu)$$

Given $\varepsilon > 0$ and a $\mu \otimes \nu$-measurable function $b(x, y)$, the *entropic optimal transport problem* is

$$\min_{\pi \in \Pi(\mu,\nu)} \mathrm{KL}(\pi | e^{-b/\varepsilon} \mu \otimes \nu), \quad \text{where } \mathrm{KL}(\pi|\bar{\pi}) = \int \log\left({}^{d\pi}/_{d\bar{\pi}}\right) d\pi \tag{25}$$

Sinkhorn algorithm/Entropic optimal transport

Let (X, μ) and (Y, ν) be two probability spaces and take the set of couplings over $\mathsf{X} \times \mathsf{Y}$ (i.e. joint laws) having marginal μ (resp. ν)

$$C = \Pi(\mu, *), \quad D = \Pi(*, \nu), \quad \Pi(\mu, \nu) = \Pi(\mu, *) \cap \Pi(*, \nu)$$

Given $\varepsilon > 0$ and a $\mu \otimes \nu$-measurable function $b(x, y)$, the *entropic optimal transport problem* is

$$\min_{\pi \in \Pi(\mu,\nu)} \mathrm{KL}(\pi | e^{-b/\varepsilon} \mu \otimes \nu), \qquad \text{where } \mathrm{KL}(\pi|\bar{\pi}) = \int \log\left({}^{d\pi}/_{d\bar{\pi}}\right) d\pi \tag{25}$$

The Sinkhorn algorithm solves (25) by initializing $\pi_0(dx, dy) = e^{-b(x,y)/\varepsilon} \mu(dx)\nu(dy)$ and by alternating "Bregman projections" onto $\Pi(\mu, *)$ and $\Pi(*, \nu)$,

$$\gamma_{n+1} = \operatorname*{argmin}_{\gamma \in \Pi(\mu,*)} \mathrm{KL}(\gamma | \pi_n), \tag{26}$$

$$\pi_{n+1} = \operatorname*{argmin}_{\pi \in \Pi(*,\nu)} \mathrm{KL}(\pi | \gamma_{n+1}). \tag{27}$$

$$\gamma_{n+1} = \operatorname*{argmin}_{\gamma \in \Pi(\mu,*)} \mathrm{KL}(\gamma|\pi_n), \tag{28}$$

$$\pi_{n+1} = \operatorname*{argmin}_{\pi \in \Pi(*,\nu)} \mathrm{KL}(\pi|\gamma_{n+1}). \tag{29}$$

The iterates of Sinkhorn (the ones above) are also given by

$$\gamma_{n+1} = \operatorname*{argmin}_{\gamma \in \Pi(\mu,*)} \mathrm{KL}(\pi_n|\gamma), \tag{30}$$

$$\pi_{n+1} = \operatorname*{argmin}_{\pi \in \Pi(*,\nu)} \mathrm{KL}(\pi|\gamma_{n+1}). \tag{31}$$

Csiszár and Tusnády show FP directly [Csiszár and Tusnády, 1984, Section 3]. Alternatively KL is a Bregman divergence and *jointly convex*, so

$$F(\pi) = \inf_{\gamma \in \Pi(\mu,*)} \Phi(\pi,\gamma) = \mathrm{KL}(p_{\mathsf{X}}\pi|\mu) \text{ is convex.} \quad \mathrm{KL}(p_{\mathsf{X}}\pi_n|\mu) \le \frac{\mathrm{KL}(\pi|\gamma_0)}{n}.$$

Natural gradient descent

Take $Y = X$ and consider the cost with u C^3, convex, with invertible Hessian

$$c(x, y) = u(y|x) = u(y) - u(x) - \langle \nabla u(x), y - x \rangle.$$

Consequently

$$-\nabla_x c(x, y) = \nabla^2 u(x)(y - x).$$

Our gradient descent thus gives

$$\begin{aligned} y_{n+1} &= x_n - \nabla^2 u(x_n)^{-1} \nabla f(x_n), \\ \nabla_x c(x_{n+1}, y_{n+1}) &= 0. \end{aligned}$$

Combining, we get natural gradient descent: $x_{n+1} - x_n = -\nabla^2 u(x_n)^{-1} \nabla f(x_n)$.

Lemma (Natural gradient descent: c-concavity and cross-convexity)

Let $f : X \to \mathbb{R}$ be twice differentiable.

i) *f is c-concave if and only if for all x, ξ,*

$$\nabla^2 f(x)(\xi,\xi) \leq \nabla^3 u(x)(\nabla^2 u(x)^{-1} \nabla f(x), \xi, \xi) + \nabla^2 u(x)(\xi,\xi); \tag{32}$$

ii) *Let $\lambda \geq 0$. f is λ-strongly c-cross-convex if and only if $f \circ \nabla u^*$ is convex, for all x, ξ,*

$$\nabla^2 f(x)(\xi,\xi) \geq \nabla^3 u(x)(\nabla^2 u(x)^{-1} \nabla f(x), \xi, \xi) + \lambda \nabla^2 u(x)(\xi,\xi). \tag{33}$$

These assumptions give new global rates for NGD as well as for Newton!

Newton

Let $Y = X$ and consider the cost

$$c(x, y) = f(y|x) = f(y) - f(x) - \langle \nabla f(x), y - x \rangle.$$

Then gradient descent with general cost reads

$$x_{n+1} - x_n = -\nabla^2 f(x_n)^{-1} \nabla f(x_n). \tag{34}$$

This is *Newton's method.* Let $0 \leq \lambda < 1$ and consider the (affine-invariant!) property:

$$0 \leq \nabla^3 f(x)((\nabla^2 f)^{-1}(x) \nabla f(x), \xi, \xi) \leq (1 - \lambda) \nabla^2 f(x)(\xi, \xi), \quad \forall x, \xi \in X. \tag{35}$$

First inequality is $f \circ \nabla f^*$ convex. This is not self-concordance (e^x vs $\log(x)$), which reads

$$|\nabla^3 f(x)(\xi, \xi, \xi)| \leq 2M \left(\nabla^2 f(x)(\xi, \xi) \right)^{3/2}, \quad \forall x, \xi \in X, \tag{36}$$

and our property gives global linear rates under (35) (for functions like e^{Ax-b}, appearing e.g. in Cominetti/San Martin (1994))