Revisiting optimization: gradient descent with a general cost

Pierre-Cyril Aubin,
joint work with Flavien Léger (INRIA)

Postdoc at INRIA Paris - SIERRA

MaLGa Seminar, May 2023
Talk based on *Gradient descent with a general cost*
available on arXiv https://arxiv.org/abs/2305.04917

Motivation: gradient descent

Take $f : \mathbb{R}^d \to \mathbb{R}$, $L > 0$ and consider gradient descent

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

For convergence of the gradient norm $\|\nabla f(x_n)\|$, we just need L-smoothness, expressed as a "descent lemma"

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

Gradient descent is just minimization of the upper bound!

To obtain (sub)linear convergence of $f(x_n)$, we need (strong) convexity to hold for a $\lambda \geq 0$

$$f(x) + \langle \nabla f(x), x' - x\rangle + \frac{\lambda}{2}\|x - x'\|^2 \leq f(x'). \tag{3}$$

How to generalize these conditions when $\|x - x'\|^2$ is "replaced" by $c(x, y)$?

Motivation: mirror descent

Take a convex $u : \mathbb{R}^d \to \mathbb{R}$ and consider its Bregman divergence

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

Assume f is smooth *relatively to u* [Bauschke et al., 2017] i.e.

$$f(x') \leq f(x) + \langle \nabla f(x), x' - x \rangle + u(x'|x). \tag{4}$$

which is equivalent to $f(x'|x) \leq u(x'|x)$.

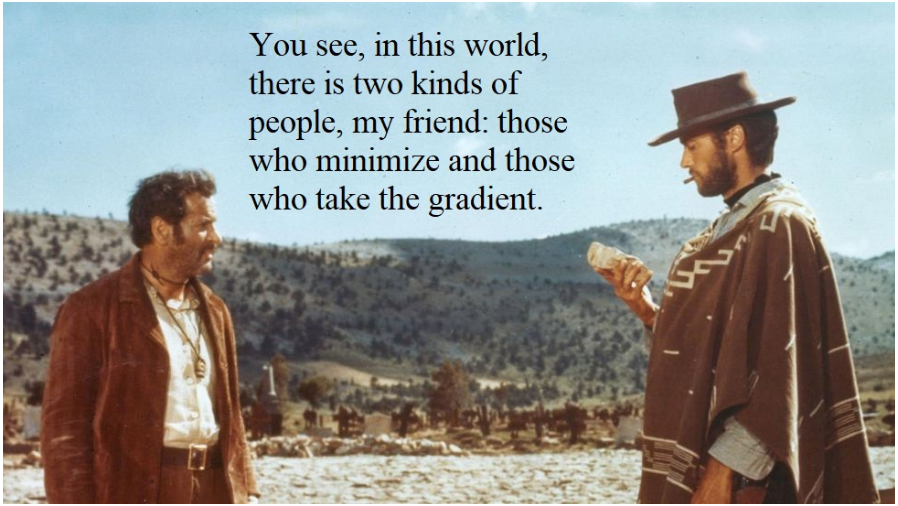
If f is also λ-strongly convex *relatively to u* [Lu et al., 2018], i.e. $f(x'|x) \geq \lambda u(x'|x)$ for $\lambda \geq 0$, we get (sub)linear convergence of $f(x_n)$ for the *mirror descent* scheme

$$x_{n+1} = \operatorname*{argmin}_{x \in X} f(x_n) + \langle \nabla f(x_n), x - x_n \rangle + u(x|x_n). \tag{5}$$

which is equivalent to

$$\nabla u(x_{n+1}) - \nabla u(x_n) = -\nabla f(x_n). \tag{6}$$

Already nice, but can we go further? To natural gradient descent and beyond?



For simplicity, we assume that minimizers exist and are unique! Otherwise we need arguments based on continuity, compactness. . . If we differentiate, then we work on open subsets of $\mathbb{R}^d$.

Executive summary: majorization–minimization

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

$$f(x) \leq \phi(x,y) \coloneqq c(x,y) + f^c(y) \coloneqq c(x,y) + \sup_{x' \in X} f(x') - c(x',y) \tag{7}$$

Do alternating minimization (AM) of the surrogate

$$y_{n+1} = \operatorname*{argmin}_{y \in Y} c(x_n, y) + f^c(y), \tag{8}$$

$$x_{n+1} = \operatorname*{argmin}_{x \in X} c(x, y_{n+1}) + f^c(y_{n+1}). \tag{9}$$

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

$$-\nabla_x c(x_n, y_{n+1}) = -\nabla f(x_n), \tag{10}$$

$$\nabla_x c(x_{n+1}, y_{n+1}) = 0. \tag{11}$$

Executive summary: convergence rates

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, for all $x, y_n \in X \times Y$,

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

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

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

These conditions extend L-smoothness and (strong) convexity when $c(x, y) = \frac{L}{2}\|x - y\|^2$

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

$$f(x_n) - f(x_*) \leq \frac{c(x_*, y_0) - c(x_0, y_0)}{n}. \tag{12}$$

Linear rates and local caracterization of c-concavity and c-cross-convexity given later.

Formal algorithm

INPUT: a set X, a point $x_0 \in X$ and a function $f : X \to \mathbb{R}$, N a number of steps
CHOOSE: a set Y and a cost $c(x, y)$
DO: For $\phi(x, y) \coloneqq c(x, y) + \sup_{x' \in X} f(x') - c(x', y)$, N steps of alternating minimization of ϕ

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

CHECK: convergence conditions for $x \in \{x_0, \dots x_N\}$, $n \in \{0 \dots N\}$

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

OUTPUT: $(x_n, y_n)_n$ iterates.

Alternating minimization

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

$$\begin{aligned} y_{n+1} &= \operatorname*{argmin}_{y \in Y} \phi(x_n, y) \\ x_{n+1} &= \operatorname*{argmin}_{x \in X} \phi(x, y_{n+1}), \end{aligned} \tag{13}$$

Inspired by [Csiszár and Tusnády, 1984], we define:

Definition (Five-point property (FPP))

We say that ϕ satisfies the FPP if for all $x \in X, y, y_0 \in Y$, with $y_0 \to x_0 \to y_1$

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

For $\lambda > 0$, ϕ has the λ-strong FPP if for all $x \in X, y, y_0 \in Y$

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

Alternating minimization - Remarks on FPP

Let $\phi(x, y) = c(x, y) + g(x) + h(y)$. Recall

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

- Five points, but actually only x, y, y_0 are free.
- Actually $0 \le \lambda < 1$ is enough, otherwise we converge in two steps for $\lambda > 1$.
- Setting $F(x) = \inf_{y \in Y} \phi(x, y)$, ($\lambda$-FP) can be written

$$F(x) \ge F(x_0) + \delta_\phi(x, y_0; x_0, y_1) + \lambda[\phi(x, y_0) - \phi(x_0, y_0)]. \tag{14}$$

  where $\delta_c(x', y'; x, y) := c(x, y') + c(x', y) - c(x, y) - c(x', y')$ is the *cross-difference* and we have $\delta_\phi = \delta_c$.
- Later we define through (λ-FP) the cross-convexity of $\phi(x, y) = c(x, y) + f^c(y)$.

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

Theorem (Convergence rates for alternating minimization)

Suppose that ϕ has a minimizer. Then:

1. *For all $n \ge 0$, $\phi(x_{n+1}, y_{n+1}) \le \phi(x_n, y_{n+1}) \le \phi(x_n, y_n)$.*
2. *If ϕ satisfies (FP). Then for any $x \in X, y \in Y$ and any $n \ge 1$,*

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

3. *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 of convergence rate

(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): After rearranging terms, (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):

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

(iii): Similarly to (ii), (λ-FP) can be written as

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

Divide both sides by $(1-\lambda)^{n+1}$ and sum from 0 to $n-1$

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

Semi-local criterion for the five-point property

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

There exists a (rather involved) semi-local characterization if $X, Y \subset \mathbb{R}^d$,

Theorem (Sufficient conditions for the five-point property)

Suppose that $\phi(x,y) = c(x,y) + g(x) + h(y)$ has a minimizer, $c \in C^4(X \times Y)$ has nonnegative cross-curvature, $\nabla^2_{xy} c(x,y)$ is everywhere invertible, X and Y have c-segments. Assume further that $F(x) = \inf_{y\in Y} \phi(x,y)$ is differentiable on X.

- *If $t \mapsto F(x(t))$ is convex on every c-segment $t \mapsto (x(t), y)$ satisfying $\nabla_x \phi(x(0), y) = 0$, then ϕ satisfies the five-point property* (FP).
- *Let $\lambda > 0$. If $t \mapsto F(x(t)) - \lambda\phi(x(t), y)$ is convex on the same c-segments as for* (i), *then ϕ satisfies the strong five-point property* (λ-FP).

Gradient descent with a general cost

Start with

$$f(x) \le c(x,y) + f^c(y) := c(x,y) + \sup_{x' \in X} f(x') - c(x',y)$$

Do alternate minimization

$$y_{n+1} = \operatorname*{argmin}_{y \in Y} c(x_n, y) + f^c(y), \tag{15}$$

$$x_{n+1} = \operatorname*{argmin}_{x \in X} c(x, y_{n+1}) + f^c(y_{n+1}). \tag{16}$$

If $f(x) = \inf_y c(x,y) + f^c(y)$ (c-concavity), then it is equivalent to

$$-\nabla_x c(x_n, y_{n+1}) = -\nabla f(x_n), \tag{17}$$

$$\nabla_x c(x_{n+1}, y_{n+1}) = 0. \tag{18}$$

Gradient descent with a general cost - Examples

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

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

1. Gradient descent: $c(x,y) = \frac{L}{2}\|x-y\|^2$ and $x_{n+1} - x_n = -\frac{1}{L}\nabla f(x_n)$.
2. Mirror descent: $c(x,y) = u(x|y)$, so $\nabla u(x_{n+1}) - \nabla u(x_n) = -\nabla f(x_n)$.
3. Natural gradient descent: $c(x,y) = u(y|x)$, so $x_{n+1} - x_n = -(\nabla^2 u(x_n))^{-1}\nabla f(x_n)$.
4. A nonlinear gradient descent: $c(x,y) = \ell(x-y)$, so $x_{n+1} - x_n = -\nabla\ell^*(\nabla f(x_n))$.
5. Riemannian gradient descent: (M, g) a Riemannian manifold. Take $X = Y = M$ and $c(x,y) = \frac{L}{2}d^2(x,y)$, so $x_{n+1} = \exp_{x_n}(-\frac{1}{L}\nabla f(x_n))$,

Cool, but what do you need to converge?
$\hookrightarrow$ Something like L-smoothness and μ-strong convexity

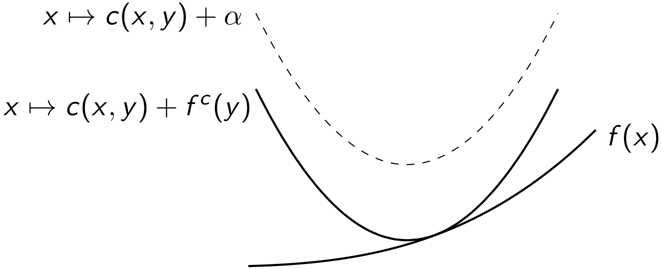
c-concavity

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 $h(y) = f^c(y) = \sup_{x' \in X} f(x') - c(x', y)$.



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

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

Definition (cross-convexity)

Suppose that f and c are differentiable. 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

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

Sketch of alternating minimization

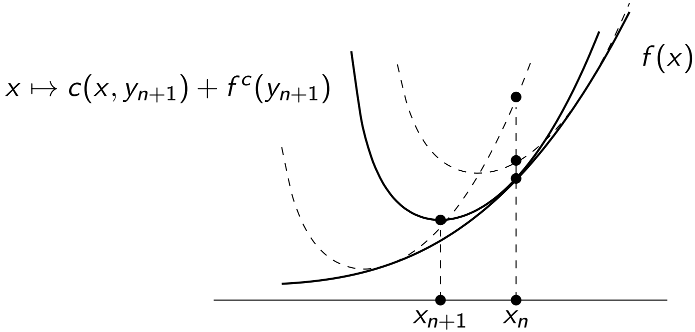


Figure: The dashed functions represent some surrogates $x \mapsto c(x, y) + f^c(y)$ for various values of y. The solid line surrogate is the one for which the value at x_n is minimized, i.e. $y = y_{n+1}$.

Let $\phi(x, y) = c(x, y) + f^c(y)$ and $\lambda \geq 0$. If f is c-concave and λ-strongly c-cross-convex then ϕ satisfies (λ-FP).

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 a twice-differentiable function. 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

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

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

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

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

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

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

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

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

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

$$f(x_n) \leq f(x) + \frac{c(x, y_0) - c(x_0, y_0)}{n}. \tag{23}$$

3. *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 $\Lambda := (1-\lambda)^{-1} > 1$*

$$f(x_n) \leq f(x) + \frac{\lambda(c(x, y_0) - c(x_0, y_0))}{\Lambda^n - 1}, \tag{24}$$

Forward–backward splitting

$$\min_{x \in X} F(x) := f(x) + g(x) \leq \phi(x, y) := c(x, y) + f^c(y) + g(x) \tag{25}$$

Additional assumption: for each $x \in X$, $\inf_{y \in Y} c(x, y) = 0$.

$$y_{n+1} = \operatorname*{argmin}_{y \in Y} c(x_n, y) + f^c(y) + g(x_n), \tag{26}$$

$$x_{n+1} = \operatorname*{argmin}_{x \in X} c(x, y_{n+1}) + f^c(y_{n+1}) + g(x). \tag{27}$$

If f is c-concave, then equivalent to

$$-\nabla_x c(x_n, y_{n+1}) = -\nabla f(x_n), \tag{28}$$

$$-\nabla_x c(x_{n+1}, y_{n+1}) = \nabla g(x_{n+1}). \tag{29}$$

Forward–backward splitting: cross-concavity

Definition (cross-concavity)

We say that a differentiable function $f\colon X \to \mathbb{R}$ is *c*-cross-concave 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

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

In addition let $\lambda > 0$. We say that f is λ-strongly *c*-cross-concave if under the same conditions as above we have

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

Caveat: f *c*-cross-concave is not in general equivalent to $(-f)$ *c*-cross-convex.

Theorem (Convergence rates for Forward–backward splitting)

Take $\bar{y}_0 \in \operatorname{argmin}_{y \in Y} c(x_0, y)$.

1. *Suppose that* f *is c-concave. Then we have the descent property*
$$f(x_{n+1}) + g(x_{n+1}) \leq f(x_n) + g(x_n).$$
2. *Suppose in addition that* f *is c-cross-convex and that* $-g$ *is c-cross-concave. Then for any* $x \in X$, $n \geq 1$,
$$f(x_n) + g(x_n) \leq f(x) + g(x) + \frac{c(x, \bar{y}_0)}{n}.$$
3. *Suppose in addition that* f *is* λ*-strongly c-cross-convex and that* $-g$ *is* μ*-strongly c-cross-concave for some* $\lambda, \mu \in [0, 1)$ *with* $\lambda + \mu > 0$. *Then for any* $x \in X$, $n \geq 1$,
$$f(x_n) + g(x_n) \leq f(x) + g(x) + \frac{(\lambda + \mu)\, c(x, \bar{y}_0)}{\Lambda^n - 1}, \quad \text{with } \Lambda = \frac{1 + \mu}{1 - \lambda}$$

Mirror descent

We take

$$c(x,y) = u(x|y) := u(x) - u(y) - \langle \nabla u(y), x - y \rangle, \tag{30}$$

We love it because

- it generalizes the square of Euclidean distances;
- it characterizes convexity, since $u(x|y) \geq 0$ iff u is convex.

Recall our scheme

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

Our gradient descent thus gives

$$\nabla u(y_{n+1}) - \nabla u(x_n) = -\nabla f(x_n),$$
$$\nabla u(x_{n+1}) = \nabla u(y_{n+1}).$$

Combining, we get mirror descent in gradient form $\nabla u(x_{n+1}) - \nabla u(x_n) = -\nabla f(x_n)$.

Definition (Relative smoothness and convexity)

Let $L > 0$, $\lambda > 0$, and consider a twice differentiable function $f : X \to \mathbb{R}$.

1. f is smooth *relatively to* u if $u - f$ is convex [Bauschke et al., 2017]. Equivalently, if $\nabla^2 f \le \nabla^2 u$, or if $f(x'|x) \le u(x'|x)$, i.e. $f(x') \le f(x) + \langle \nabla f(x), x' - x\rangle + u(x'|x)$.
2. f is λ-strongly convex *relatively to* u [Lu et al., 2018] if $f - \lambda u$ is convex. Equivalently, if $\nabla^2 f \ge \lambda \nabla^2 u$, or if $f(x'|x) \ge \lambda u(x'|x)$.

Naturally we want to minimize the upperbound given 1.:

$$x_{n+1} = \operatorname*{argmin}_{x \in X} \tilde{\phi}(x, x_n) = f(x_n) + \langle \nabla f(x_n), x - x_n\rangle + u(x|x_n) = f(x) + (u - f)(x|x_n). \quad (31)$$

Buy we can also do

$$\phi(x, y) = u(x|y) + f^c(y).$$

Actually we have $\tilde{\phi}(x, \tilde{y}) = \phi(x, y)$ when $\nabla u(y) = \nabla u(\tilde{y}) - \nabla f(\tilde{y})$ (just a reparameterization).

Mirror descent: c-concavity and cross-convexity

Proposition (c-concavity is relative smoothness)

Suppose that ∇u is surjective as a map from X to X^*. Then f is c-concave for $c(x,y)=u(x|y)$ if and only if f is smooth relatively to u.

Proposition (cross-convexity is convexity)

Take $c(x,y)=u(x|y)$. Then f is c-cross-convex if and only if f is convex. More generally, let $\lambda>0$. Then f is λ-strongly c-cross-convex if and only if f is λ-strongly convex relatively to u.

We recover the classical convergence rates:

- sublinear when f is convex and smooth relatively to u [Bauschke et al., 2017]
- linear if in addition f is λ-strongly convex relatively to u [Lu et al., 2018].

Natural gradient descent

Take $Y = X$ and consider the cost

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

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.

1. *f is c-concave if and only if for all x, ξ,*

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

2. *Let $\lambda \ge 0$. f is λ-strongly c-cross-convex if and only if, for all x, ξ,*

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

These assumptions give new global rates for NGD!

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*. The smoothness and convexity assumptions on f can be combined as follows. Let $0 \leq \lambda < 1$ and consider the (affine-invariant!) property: for all x, ξ,

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

This is not self-concordance (check e^x and $\log(x)$), i.e.

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

and our property gives global rates (which self-concordance doesn't)!

Riemannian gradient descent

For $c(x, y) = \frac{L}{2} d^2(x, y)$ on a manifold M away from the cut locus, the relation $\xi = -\nabla_x c(x, y)$ defines a tangent vector $\xi \in T_x M$, i.e. for exp the (Riemannian) exponential map

$$y = \exp_x(\xi / L).$$

We obtain as before $x_{n+1} = \exp_{x_n}\Big(-\frac{1}{L} \nabla f(x_n)\Big)$.

Proposition

Let $c(x, y) = \frac{L}{2} d^2(x, y)$. Suppose that (M, g) has nonnegative sectional curvature. Then

1. *f geodesically convex $\implies$ f c-cross-convex.*
2. *$-g$ c-cross-concave $\implies$ g geodesically convex.*

Suppose that (M, g) has nonpositive sectional curvature. Then

1. *f c-cross-convex $\implies$ f geodesically convex.*
2. *g geodesically convex $\implies$ $-g$ c-cross-concave.*

Riemannian gradient descent

1. f is c-concave;
2. f has L-Lipschitz gradients;
3. $\nabla^2 f \leq L\mathrm{g}$;
4. $f(x) \leq f(\bar{x}) + \langle \nabla f(\bar{x}), \xi \rangle + \frac{L}{2} d^2(x, \bar{x})$, where $x = \exp_{\bar{x}}(\xi)$.

Proposition

The following statements hold.

- $3 \iff 4$
- *Suppose that* (M, g) *has nonnegative curvature. Then* $1 \implies 3$.
- *Suppose that* (M, g) *has nonpositive curvature. Then* $3 \implies 1$.
- $2 \implies 3$

POCS (Projection Onto Convex Sets)

Let $(H, \|\cdot\|)$ be a Euclidean space and let B, C be two closed convex subsets of H. The POCS algorithm, see [Bauschke and Combettes, 2011], searches for $B \cap C$ by successive projections onto B and C: given $x_n \in B$, compute

$$\begin{aligned} y_{n+1} &= \operatorname*{argmin}_{y \in C} \|x_n - y\|, \\ x_{n+1} &= \operatorname*{argmin}_{x \in B} \|x - y_{n+1}\|. \end{aligned} \tag{37}$$

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

1. Take $X = Y = H$, with the cost $c(x, y) = \frac{1}{2}\|x - y\|^2$ and the indicator functions $g = \iota_B$ and $h = \iota_C$, set $\phi(x, y) = c(x, y) + g(x) + h(y)$.
2. Take $X = B$, $Y = C$ and consider the function $\phi(x, y) = \frac{1}{2}\|x - y\|^2$.

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

Sinkhorn algorithm/Entropic optimal transport

Let (X, μ) and (Y, ν) be two probability spaces and take the set of couplings over $X \times Y$ (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

$$\min_{\pi \in \Pi(\mu,\nu)} \mathrm{KL}(\pi | e^{-b/\varepsilon} \mu \otimes \nu), \qquad \text{where } \mathrm{KL}(\pi|\bar{\pi}) = \int \log\left(d\pi/d\bar{\pi}\right) d\pi \tag{38}$$

The Sinkhorn algorithm solves (38) by initializing $\pi_0(dx, dy) = e^{-b(x,y)/\varepsilon} \mu(dx)\nu(dy)$ and by alternating "Bregman projections" onto $\Pi(\mu, *)$ and $\Pi(*, \nu)$,

$$\gamma_{n+1} = \operatorname*{argmin}_{\gamma \in \Pi(\mu,*)} \mathrm{KL}(\gamma|\pi_n), \tag{39}$$

$$\pi_{n+1} = \operatorname*{argmin}_{\pi \in \Pi(*,\nu)} \mathrm{KL}(\pi|\gamma_{n+1}). \tag{40}$$

$$\gamma_{n+1} = \operatorname*{argmin}_{\gamma \in \Pi(\mu,*)} \mathrm{KL}(\gamma|\pi_n), \tag{41}$$

$$\pi_{n+1} = \operatorname*{argmin}_{\pi \in \Pi(*,\nu)} \mathrm{KL}(\pi|\gamma_{n+1}). \tag{42}$$

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

$$\gamma_{n+1} = \operatorname*{argmin}_{\gamma \in \Pi(\mu,*)} \mathrm{KL}(\pi_n|\gamma), \tag{43}$$

$$\pi_{n+1} = \operatorname*{argmin}_{\pi \in \Pi(*,\nu)} \mathrm{KL}(\pi|\gamma_{n+1}). \tag{44}$$

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

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

Expectation–Maximization (EM)

Let X be a set of observed data, Z be a latent space and let $\{p_\theta \in \mathcal{P}(\mathsf{X} \times \mathsf{Z}) : \theta \in \Theta\}$ be a *statistical model*, where Θ is a set of parameters. Having observed $\mu \in \mathcal{P}(\mathsf{X})$ we want to find $\theta \in \Theta$ that maximizes the *likelihood*. This is equivalent to

$$\min_{\theta \in \Theta} F(\theta) = \mathrm{KL}(\mu | p_{\mathsf{X}} p_\theta), \tag{45}$$

We use the *data processing inequality*

$$F(\theta) = \mathrm{KL}(\mu | p_{\mathsf{X}} p_\theta) \leq \mathrm{KL}(\pi | p_\theta) =: \Phi(\theta, \pi), \tag{46}$$

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])

Conclusion: What have we seen?

To minimize f on a set X, we chose a set Y and a cost $c(x,y)$.
For $\phi(x,y) := c(x,y) + \sup_{x'\in X} f(x') - c(x',y)$, we did alternating minimization of ϕ

$$y_{n+1} = \operatorname*{argmin}_{y\in Y} \phi(x_n, y)$$

$$x_{n+1} = \operatorname*{argmin}_{x\in X} \phi(x, y_{n+1}).$$

We also did a forward–backward version of this and covered MD/NGD/RGD/Sinkhorn/EM...
We have seen that (sub)linear rates could be obtained based on

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

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

Tell me about your favorite algorithm and we can see if it is an alternating minimization!

Thank you for your attention!

Other interests of mine: backward SDEs+optimal control (V. de Bortoli), kernels+ mean field control (A. Bensoussan)

References I

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., Bolte, J., and Teboulle, M. (2017).
A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications.
Math. Oper. Res., 42(2):330–348.

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*.

Lu, H., Freund, R. M., and Nesterov, Y. (2018).
Relatively smooth convex optimization by first-order methods, and applications.
SIAM J. Optim., 28(1):333–354.

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.