Alternating minimization and gradient descent with $c(x, y)$ cost

Pierre-Cyril Aubin-Frankowski

Postdoc at TU Wien - VADOR

One World Optimization Seminar in Vienna, June 2024

joint work with Flavien Léger (INRIA Paris)

Motivation: extending gradient descent

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

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

To have $\|\nabla f(x_n)\| \overset{n \to \infty}{\longrightarrow} 0$, L-smoothness ($\nabla^2 f \leq L \,\mathrm{Id}$) suffices, reading 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!

Motivation: extending gradient descent

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

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

To have $\|\nabla f(x_n)\| \overset{n \to \infty}{\longrightarrow} 0$, L-smoothness ($\nabla^2 f \leq L\,\mathrm{Id}$) suffices, reading 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 $\lambda \geq 0$

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

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

Motivation: extending gradient descent

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

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

To have $\|\nabla f(x_n)\| \xrightarrow{n\to\infty} 0$, L-smoothness ($\nabla^2 f \leq L\,\mathrm{Id}$) suffices, reading 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 $\lambda \geq 0$

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

There are three objects: i) an algorithm; ii) a regularizer; iii) a class of functions
How to generalize this setting when $\|x - x'\|^2$ is "replaced" by $c(x, y)$?

Systematic majorization–minimization with a cost

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

$$f(x) \leq \phi(x, y) \coloneqq c(x, y) + f^c(y) \coloneqq c(x, y) + \sup_{x' \in X} f(x') - c(x', y) \tag{4}$$

Do alternating minimization (AM) of the surrogate

$$y_{n+1} = \operatorname*{argmin}_{y \in Y} c(x_n, y) + f^c(y), \tag{5}$$

$$x_{n+1} = \operatorname*{argmin}_{x \in X} c(x, y_{n+1}) + f^c(y_{n+1}). \tag{6}$$

Systematic majorization–minimization with a cost

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

$$f(x) \leq \phi(x,y) \coloneqq c(x,y) + f^c(y) \coloneqq c(x,y) + \sup_{x' \in X} f(x') - c(x',y) \tag{4}$$

Do alternating minimization (AM) of the surrogate

$$y_{n+1} = \operatorname*{argmin}_{y \in Y} c(x_n, y) + f^c(y), \tag{5}$$

$$x_{n+1} = \operatorname*{argmin}_{x \in X} c(x, y_{n+1}) + f^c(y_{n+1}). \tag{6}$$

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 $\nabla f(x) = \nabla_1 \phi(x, \bar{y}(x))$)

$$-\nabla_x c(x_n, y_{n+1}) = -\nabla f(x_n), \tag{7}$$

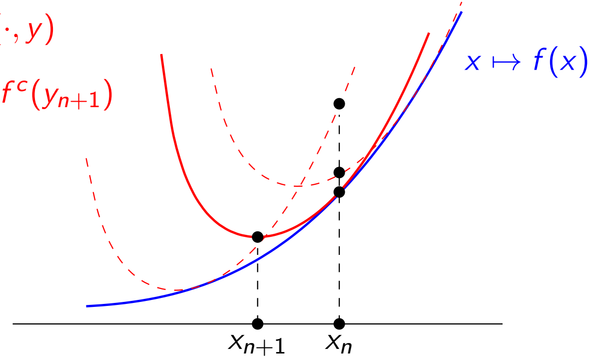
$$\nabla_x c(x_{n+1}, y_{n+1}) = 0. \tag{8}$$

For a quadratic c, we recover gradient descent!

Visual sketch of alternating minimization

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

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

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

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

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 $f^c(y_{n+1}) = f(x_n) - c(x_n, y_{n+1})$,

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

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 $f^c(y_{n+1}) = f(x_n) - c(x_n, y_{n+1})$,

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

For $c(x, y) = \frac{L}{2}\|x - y\|^2$, we get

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

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

$$f(x_n) - f(x_*) \leq \frac{c(x_*, y_0) - c(x_0, y_0)}{n}. \tag{9}$$

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

What are we going to see?

1. Motivation
2. Alternating minimization and GradDesc with GenCost
3. c-concavity and c-cross-convexity
4. Examples

Alternating minimization (AM)

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

$$\begin{aligned} y_{n+1} &= \operatorname*{argmin}_{y \in Y} \phi(x_n, y) \\ x_{n+1} &= \operatorname*{argmin}_{x \in X} \phi(x, y_{n+1}), \end{aligned} \tag{10}$$

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

Alternating minimization (AM)

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

$$\begin{aligned} y_{n+1} &= \operatorname*{argmin}_{y \in Y} \phi(x_n, y) \\ x_{n+1} &= \operatorname*{argmin}_{x \in X} \phi(x, y_{n+1}), \end{aligned} \tag{10}$$

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

"Paradoxically, the apparent lack of sophistication may also account for the unpopularity [of block coordinate descent] as a subject for investigation by optimization researchers, who have usually been quick to suggest alternative approaches in any given situation."

Coordinate Descent Algorithms, Stephen J. Wright, MathProg B, 2015

Alternating minimization (AM)

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

$$\begin{aligned} y_{n+1} &= \operatorname*{argmin}_{y \in Y} \phi(x_n, y) \\ x_{n+1} &= \operatorname*{argmin}_{x \in X} \phi(x, y_{n+1}), \end{aligned} \tag{10}$$

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

"Paradoxically, the apparent lack of sophistication may also account for the unpopularity [of block coordinate descent] as a subject for investigation by optimization researchers, who have usually been quick to suggest alternative approaches in any given situation."

Coordinate Descent Algorithms, Stephen J. Wright, MathProg B, 2015

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

Some results for AM exist, but based on L-smoothness or convexity
[Beck and Tetruashvili, 2013, Beck, 2015] or prox and KL-inequality [Attouch et al., 2010]

Alternating minimization (AM)

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

$$\begin{aligned} y_{n+1} &= \operatorname*{argmin}_{y \in Y} \phi(x_n, y) \\ x_{n+1} &= \operatorname*{argmin}_{x \in X} \phi(x, y_{n+1}), \end{aligned} \tag{11}$$

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

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

Definition (Five-point property (FPP))

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

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

Note that (λ-FP) forces that $y_0 \to x_0 \to y_1$ as in (11).

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

Theorem (Convergence rates for alternating minimization)

Suppose that ϕ has a minimizer. Then:

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

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

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

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

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

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

Proof of convergence rate

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

(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 of convergence rate

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

(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):

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

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

(Forward-Backward) Gradient descent with a general cost

Start with

$$f(x) + g(x) \leq \phi(x, y) := g(x) + c(x, y) + \sup_{x' \in X} f(x') - c(x', y)$$

Do alternate minimization

$$y_{n+1} = \operatorname*{argmin}_{y \in Y} c(x_n, y) + f^c(y) + g(x_n), \tag{12}$$

$$x_{n+1} = \operatorname*{argmin}_{x \in X} c(x, y_{n+1}) + f^c(y_{n+1}) + g(x). \tag{13}$$

Let $F(x) = \inf_y \phi(x, y)$ (c-concavity is $f = F$), and assume we are allowed to differentiate, then

$$-\nabla_x c(x_n, y_{n+1}) = -\nabla F(x_n), \tag{14}$$

$$\nabla_x c(x_{n+1}, y_{n+1}) = \nabla g(x_{n+1}). \tag{15}$$

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)

i) Gradient descent: $c(x, y) = \frac{L}{2}\|x - y\|^2$ and $x_{n+1} - x_n = -\frac{1}{L}\nabla f(x_n)$.

ii) Mirror descent: $c(x, y) = u(x|y)$, so $\nabla u(x_{n+1}) - \nabla u(x_n) = -\nabla f(x_n)$.

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)

i) Gradient descent: $c(x, y) = \frac{L}{2}\|x - y\|^2$ and $x_{n+1} - x_n = -\frac{1}{L}\nabla f(x_n)$.

ii) Mirror descent: $c(x, y) = u(x|y)$, so $\nabla u(x_{n+1}) - \nabla u(x_n) = -\nabla f(x_n)$.

iii) Natural gradient descent: $c(x, y) = u(y|x)$, so $x_{n+1} - x_n = -(\nabla^2 u(x_n))^{-1}\nabla f(x_n)$.

iv) A nonlinear gradient descent: $c(x, y) = \ell(x - y)$, so $x_{n+1} - x_n = -\nabla\ell^*(\nabla f(x_n))$.

v) Riemannian gradient descent: (M, g) a Riemannian manifold. Take $X = Y = M$ and $c(x, y) = \frac{L}{2}d^2(x, y)$, so $x_{n+1} = \exp_{x_n}(-\frac{1}{L}\nabla f(x_n))$,

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

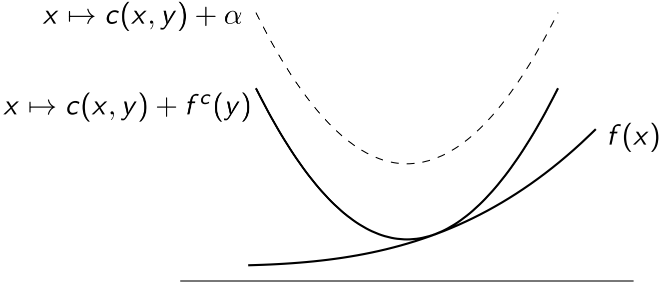
c-concavity

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

$$f(x) = \inf_{y \in Y} c(x,y) + h(y), \tag{16}$$

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)

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

$$f(x) \geq f(\bar{x}) + \delta_c(x, \bar{y}; \bar{x}, \hat{y}). \tag{17}$$

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

$$f(x) \geq f(\bar{x}) + \delta_c(x, \bar{y}; \bar{x}, \hat{y}) + \lambda(c(x, \bar{y}) - c(\bar{x}, \bar{y})). \tag{18}$$

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

$$\nabla^2 f(\bar{x}) \geq \nabla^2_{xx} c(\bar{x}, \hat{y}) - \nabla^2_{xx} c(\bar{x}, \bar{y}). \tag{19}$$

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

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 $x \in X, n \geq 1$,

$$f(x_n) \leq f(x) + \frac{c(x, y_0) - c(x_0, y_0)}{n}. \tag{20}$$

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

$$f(x_n) \leq f(x) + \frac{\lambda(c(x, y_0) - c(x_0, y_0))}{\Lambda^n - 1}, \tag{21}$$

Forward-backward is also possible. But now, on to examples!

Mirror descent

For $u : X \to \mathbb{R}$ differentiable, consider

$$c(x, y) = u(x|y) \coloneqq u(x) - u(y) - \langle \nabla u(y), x - y \rangle, \tag{22}$$

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

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

Our gradient descent thus gives

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

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 $f \ C^2$.

i) f is smooth *relatively to* u if $u - f$ is convex [Bauschke et al., 2017]. Equivalently, if $\nabla^2 f \leq \nabla^2 u$, or if $f(x'|x) \leq u(x'|x)$, i.e. $f(x') \leq f(x) + \langle \nabla f(x), x' - x \rangle + u(x'|x)$.

ii) f is λ-strongly convex *relatively to* u [Lu et al., 2018] if $f - \lambda u$ is convex. Equivalently, if $\nabla^2 f \geq \lambda \nabla^2 u$, or if $f(x'|x) \geq \lambda u(x'|x)$.

Naturally we want to minimize the upperbound given by 1.:

$$x_{n+1} = \operatorname*{argmin}_{x \in X} \tilde{\phi}(x, x_n) = f(x_n) + \langle \nabla f(x_n), x - x_n \rangle + u(x|x_n) = f(x) + (u - f)(x|x_n). \quad (23)$$

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].

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

- *i) f geodesically convex $\implies$ f c-cross-convex.*
- *ii) $-g$ c-cross-concave $\implies$ g geodesically convex.*

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

- *i) f c-cross-convex $\implies$ f geodesically convex.*
- *ii) g geodesically convex $\implies$ $-g$ c-cross-concave.*

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,

$$\nabla^2 f(x)(\xi,\xi) \leq \nabla^3 u(x)(\nabla^2 u(x)^{-1}\nabla f(x), \xi, \xi) + \nabla^2 u(x)(\xi,\xi); \tag{24}$$

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,

$$\nabla^2 f(x)(\xi,\xi) \geq \nabla^3 u(x)(\nabla^2 u(x)^{-1}\nabla f(x), \xi, \xi) + \lambda \nabla^2 u(x)(\xi,\xi). \tag{25}$$

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

$$x_{n+1} - x_n = -\nabla^2 f(x_n)^{-1} \nabla f(x_n). \tag{26}$$

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

$$0 \leq \nabla^3 f(x)((\nabla^2 f)^{-1}(x)\nabla f(x), \xi, \xi) \leq (1-\lambda)\nabla^2 f(x)(\xi,\xi), \quad \forall x, \xi \in X. \tag{27}$$

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

$$|\nabla^3 f(x)(\xi,\xi,\xi)| \leq 2M \left(\nabla^2 f(x)(\xi,\xi)\right)^{3/2}, \quad \forall x, \xi \in X, \tag{28}$$

and our property gives global linear rates under (27) (for functions like e^{Ax-b}, appearing e.g. in Cominetti/San Martin (1994))

Riemannian gradient descent

i) f is c-concave;

ii) f has L-Lipschitz gradients;

iii) $\nabla^2 f \leq L\mathrm{g}$;

iv) $f(x) \leq f(\bar{x}) + \langle \nabla f(\bar{x}), \xi \rangle + \frac{L}{2} d^2(x, \bar{x})$, where $x = \exp_{\bar{x}}(\xi)$.

Proposition

The following statements hold.

- *iii)* $\iff$ *iv)*
- *Suppose that* (M, g) *has nonnegative curvature. Then i)* $\implies$ *iii).*
- *Suppose that* (M, g) *has nonpositive curvature. Then iii)* $\implies$ *i).*
- *ii)* $\implies$ *iii)*

Conclusion: What is to be seen in the paper?

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

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

There is a forward–backward version of this and we cover MD/NGD/RGD/Sinkhorn/EM...
(Sub)linear rates can be obtained based on upper/lower bounds

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

c-concavity for revisiting optimization algorithms!

c-concavity and c-cross-convexity generalize smoothness and convexity and encompass many algorithms! New assumptions for global convergence of natural gradient descent/Newton.

Conclusion: What is to be seen in the paper?

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

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

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

There is a … EM. . .
(Sub)linea …

Thank you for your attention!

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

c-concavity for revisiting optimization algorithms!

c-concavity and c-cross-convexity generalize smoothness and convexity and encompass many algorithms! New assumptions for global convergence of natural gradient descent/Newton.

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 $x_n \in B$,

$$\begin{aligned} y_{n+1} &= \underset{y \in C}{\operatorname{argmin}} \|x_n - y\|, \\ x_{n+1} &= \underset{x \in B}{\operatorname{argmin}} \|x - y_{n+1}\|. \end{aligned} \tag{29}$$

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 $x_n \in B$,

$$\begin{aligned} y_{n+1} &= \underset{y \in C}{\operatorname{argmin}} \|x_n - y\|, \\ x_{n+1} &= \underset{x \in B}{\operatorname{argmin}} \|x - y_{n+1}\|. \end{aligned} \tag{29}$$

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*,

$$\min_{\theta \in \Theta} F(\theta) = \mathrm{KL}(\mu | p_{\mathsf{X}} p_\theta), \tag{30}$$

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

$$\min_{\pi \in \Pi(\mu,\nu)} \mathrm{KL}(\pi | e^{-b/\varepsilon} \mu \otimes \nu), \quad \text{where } \mathrm{KL}(\pi|\bar{\pi}) = \int \log \left(d\pi / d\bar{\pi} \right) d\pi \tag{31}$$

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

$$\min_{\pi \in \Pi(\mu,\nu)} \mathrm{KL}(\pi | e^{-b/\varepsilon} \mu \otimes \nu), \qquad \text{where } \mathrm{KL}(\pi|\bar{\pi}) = \int \log\left(d\pi/d\bar{\pi}\right) d\pi \tag{31}$$

The Sinkhorn algorithm solves (31) by initializing $\pi_0(dx, dy) = e^{-b(x,y)/\varepsilon} \mu(dx)\nu(dy)$ and by alternating "Bregman projections" onto $\Pi(\mu, *)$ and $\Pi(*, \nu)$,

$$\gamma_{n+1} = \operatorname*{argmin}_{\gamma \in \Pi(\mu,*)} \mathrm{KL}(\gamma | \pi_n), \tag{32}$$

$$\pi_{n+1} = \operatorname*{argmin}_{\pi \in \Pi(*,\nu)} \mathrm{KL}(\pi | \gamma_{n+1}). \tag{33}$$

$$\gamma_{n+1} = \operatorname*{argmin}_{\gamma \in \Pi(\mu,*)} \mathrm{KL}(\gamma|\pi_n), \tag{34}$$

$$\pi_{n+1} = \operatorname*{argmin}_{\pi \in \Pi(*,\nu)} \mathrm{KL}(\pi|\gamma_{n+1}). \tag{35}$$

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

$$\gamma_{n+1} = \operatorname*{argmin}_{\gamma \in \Pi(\mu,*)} \mathrm{KL}(\pi_n|\gamma), \tag{36}$$

$$\pi_{n+1} = \operatorname*{argmin}_{\pi \in \Pi(*,\nu)} \mathrm{KL}(\pi|\gamma_{n+1}). \tag{37}$$

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

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

References I

Attouch, H., Bolte, J., Redont, P., and Soubeyran, A. (2010).
Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality.
Mathematics of Operations Research, 35(2):438–457.

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.

References II

Beck, A. (2015).
On the convergence of alternating minimization for convex programming with applications to iteratively reweighted least squares and decomposition schemes.
SIAM Journal on Optimization, 25(1):185–209.

Beck, A. and Tetruashvili, L. (2013).
On the convergence of block coordinate descent type methods.
SIAM Journal on Optimization, 23(4):2037–2060.

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

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*.

References III

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.