Gradient descent with a general cost

Flavien Léger

Inria

joint works with Pierre-Cyril Aubin-Frankowski

Outline

1. A new family of algorithms

Gradient descent as alternating minimization

General method unifies gradient/mirror/natural gradient/Riemannian descent

2. Convergence theory

Generalized smoothness and convexity

Optimal transport theory → local characterizations

3. Applications

Global rates for Newton

Explicit vs. implicit Riemannian gradient descent

1. Gradient descent as minimizing movement

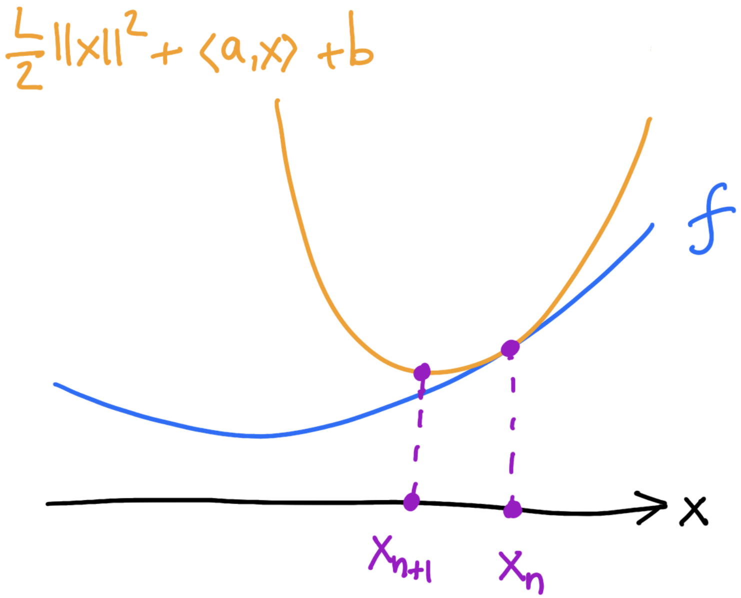
$$x_{n+1} = x_n - \frac{1}{L}\nabla f(x_n),$$

objective function $f\colon \mathbb{R}^d \to \mathbb{R}$

DEFINITION

f is L-smooth if

$$\nabla^2 f \preceq L I_{d\times d}$$



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

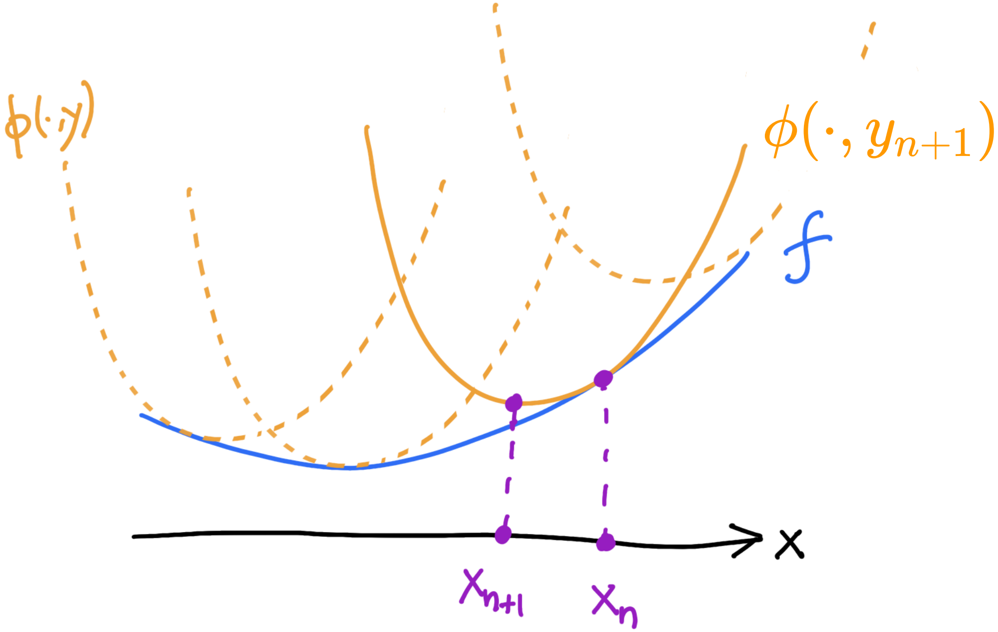
Two steps:

1) majorize: find the tangent parabola (“surrogate”)

2) minimize: minimize the surrogate

Reformulating the majorize step

Family of majorizing functions $\phi(x, y)$



Majorize step $\leftrightarrow$ y-update:

$$y_{n+1} = \arg\min_{y} \phi(x_n, y)$$

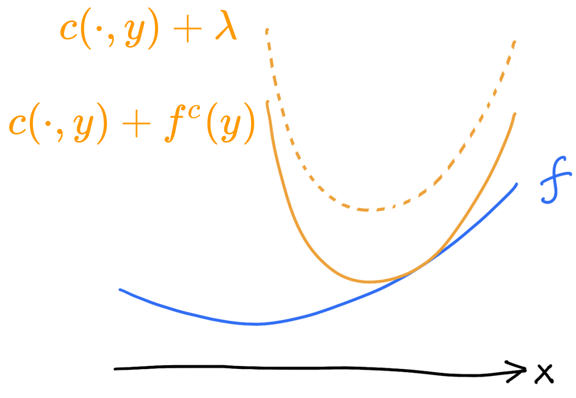
Minimize step $\leftrightarrow$ x-update:

$$x_{n+1} = \arg\min_{x} \phi(x, y_{n+1})$$

General cost (Moreau '66)

Given: X and $f: X \to \mathbb{R}$

Choose: Y and $c(x,y)$



$$f(x) \le \underbrace{c(x,y) + f^c(y)}_{\phi(x,y)}$$

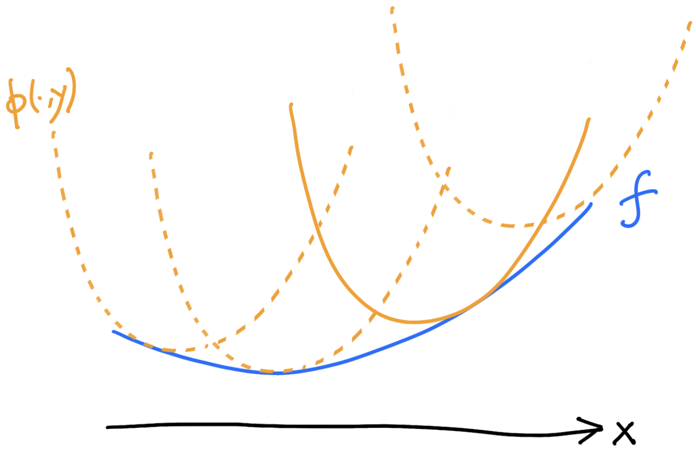
DEFINITION *c*-transform

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

DEFINITION

f is c-concave if

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



c-concavity is smoothness

DEFINITION

f is c-concave if

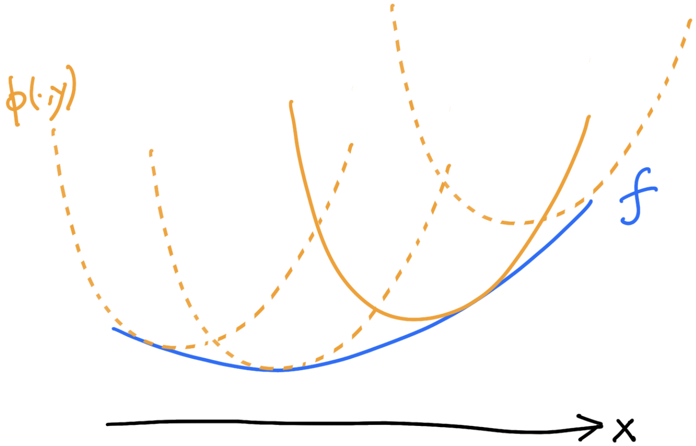
$$f(x) = \inf_{y \in Y} c(x,y) + f^c(y)$$

Example

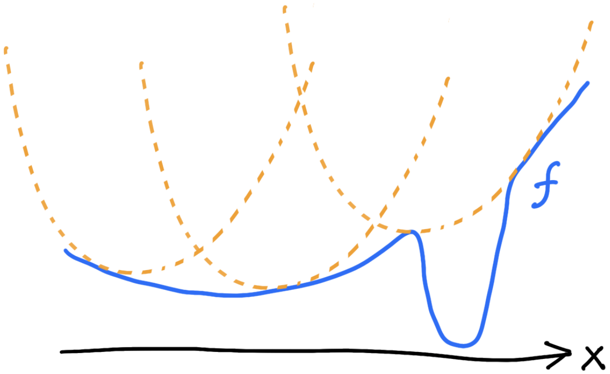
$c(x,y) = \frac{L}{2}\|x-y\|^2$

f is c-concave $\iff \nabla^2 f \le L I_{d\times d}$

$$\inf_x f(x) = \inf_{x,y} c(x,y) + f^c(y)$$



f is c-concave

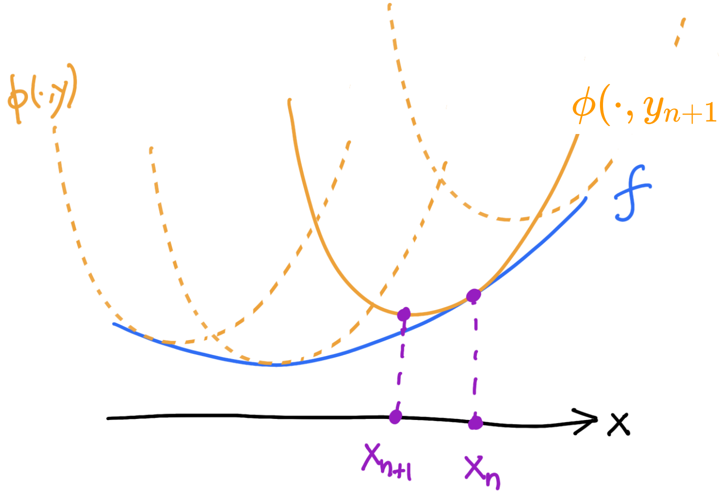


f is not c-concave

Gradient descent with a general cost

(FL–PCAF '23)

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



ALGORITHM

“majorize” $$y_{n+1} = \arg\min_{y \in Y} c(x_n, y) + f^c(y)$$

“minimize” $$x_{n+1} = \arg\min_{x \in X} c(x, y_{n+1}) + f^c(y_{n+1})$$

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

$$-\nabla_x c(x_n, y_{n+1}) = -\nabla f(x_n)$$
$$\Updownarrow$$
$$y_{n+1} = \text{c-exp}_{x_n}(-\nabla f(x_n))$$

Some examples

$$c(x,y) = \underbrace{u(x) - u(y) - \langle \nabla u(y), x - y\rangle}_{=:\, u(x|y)} \longrightarrow$$ *mirror descent*

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

$$c(x,y) = u(y|x) \longrightarrow$$ *natural gradient descent*

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

Newton

$$c(x,y) = \tfrac{L}{2} d_M^2(x,y) \longrightarrow$$ *Riemannian gradient descent*

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

1. A new family of algorithms

Gradient descent as alternating minimization

General method unifies gradient/mirror/natural gradient/Riemannian descent

2. Convergence theory

Generalized smoothness and convexity

Optimal transport theory → local characterizations

3. Applications

Global rates for Newton

Explicit vs. implicit Riemannian gradient descent

2. Cross-convexity

Cross-difference: $\delta_c(x', y'; x, y) = [c(x, y') + c(x', y)] - [c(x, y) + c(x', y')]$

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

DEFINITION

f is λ-strongly c-cross-convex if for all x, x_n,

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

Example: $c(x, y) = \frac{L}{2}\|x - y\|^2$

$$f(x) \geq f(x_n) + \langle \nabla f(x_n), x - x_n \rangle + \frac{\lambda L}{2}\|x - x_n\|^2$$

Convergence rates

THEOREM (FL–PCAF '23)

If f is c-concave and c-cross-convex then

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

If f is λ-strongly c-cross-convex with $0 < \lambda < 1$, then

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

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

Proof.

("Fenchel–Young inequality")

$$\left.\begin{aligned} f(x_{n+1}) &\le c(x_{n+1},y_{n+1}) + f^c(y_{n+1}) \\ f(x_n) &= c(x_n,y_{n+1}) + f^c(y_{n+1}) \end{aligned}\right\} \implies f(x_{n+1}) \le f(x_n) - [c(x_n,y_{n+1}) - c(x_{n+1},y_{n+1})]$$

(c-concavity)

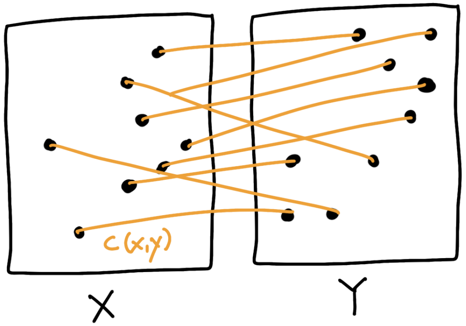
$$f(x_n) \le f(x) + c(x,y_n) - c(x,y_{n+1})$$

(cross-convexity)

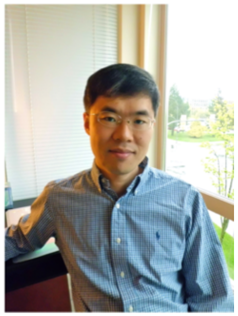
$$+c(x_n,y_{n+1}) - c(x_n,y_n)$$

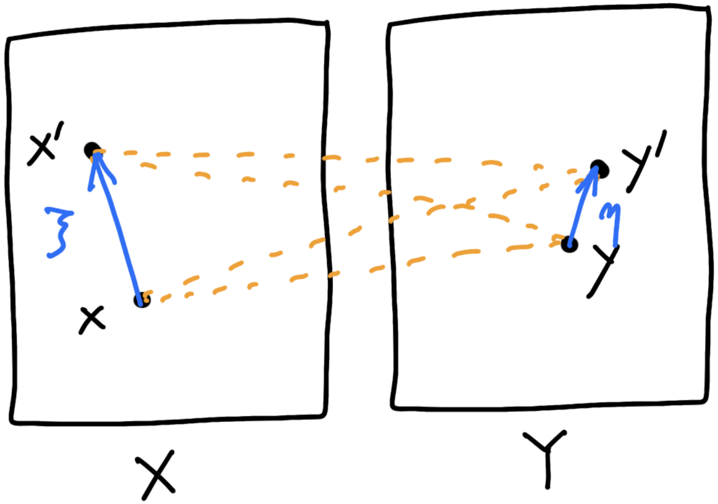
$$\implies f(x_{n+1}) \le f(x) + [c(x,y_n) - c(x_n,y_n)] - [c(x,y_{n+1}) - c(x_{n+1},y_{n+1})]$$

The Kim–McCann geometry



$$\inf_{\pi \in \Pi(\mu,\nu)} \iint_{X\times Y} c(x,y)\,\pi(dx,dy)$$





$$\delta_c(x',y';x,y) =$$
$$[c(x,y') + c(x',y)] - [c(x,y) + c(x',y')$$

$$\delta_c(x+\xi, y+\eta; x, y) = \underbrace{-\nabla^2_{xy} c(x,y)(\xi,\eta)}_{\text{Kim–McCann metric ('10)}} + o(|\xi|^2 + |\eta|^2)$$

➡ Kim–McCann geodesics

➡ Kim–McCann curvature: ***cross-curvature***

Cross-curvature

DEFINITION (Ma–Trudinger–Wang '05)

The cross-curvature or Ma–Trudinger–Wang tensor is

$$\mathfrak{S}_c(\xi, \eta) = (c_{ik\bar{s}} c^{\bar{s}t} c_{t\bar{\jmath}\bar{\ell}} - c_{i\bar{\jmath}k\bar{\ell}}) \xi^i \eta^{\bar{\jmath}} \xi^k \eta^{\bar{\ell}}$$

$$c_{i\bar{\jmath}} = \frac{\partial^2 c}{\partial x^i \partial y^{\bar{\jmath}}}, \ldots$$

THEOREM (Kim–McCann '11)

$$\mathfrak{S}_c \geq 0 \iff c(x(t), y) - c(x(t), y') \text{ convex in } t$$

for any Kim–McCann geodesic $t \mapsto (x(t), y)$

A local criteria for cross-convexity

Suppose that c has nonnegative cross-curvature.

THEOREM (Trudinger–Wang '06)

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.

THEOREM (FL–PCAF '23)

Let $\lambda > 0$. Suppose that

$$t \mapsto f(x(t)) - \lambda c(x(t), \bar{y})$$

is convex on every Kim–McCann geodesic $t \mapsto (x(t), \bar{y})$ satisfying $\nabla_x c(x(0), \bar{y}) = 0$. Then f is λ-strongly c-cross-convex.

1. A new family of algorithms

Gradient descent as alternating minimization

General method unifies gradient/mirror/natural gradient/Riemannian descent

2. Convergence theory

Generalized smoothness and convexity

Optimal transport theory → local characterizations

3. Applications

Global rates for Newton

Explicit vs. implicit Riemannian gradient descent

Global rates for Newton's method

$c(x,y)=u(y|x) \longrightarrow$ *Natural gradient descent:*

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

THEOREM (FL–PCAF '23)

If

$$\nabla^3 u(\nabla^2 u^{-1}\nabla f,-,-)\leq \nabla^2 f\leq \nabla^2 u+\nabla^3 u(\nabla^2 u^{-1}\nabla f,-,-)$$

then

$$f(x_n)\leq f(x)+\frac{u(x_0|x)}{n}$$

Newton's method: new global convergence rate.

New condition on f similar but different from self-concordance

Explicit vs. implicit Riemannian

$$\operatorname*{minimize}_{x\in M} f(x) \qquad\qquad c(x,y) = \frac{1}{2\tau} d_M^2(x,y)$$

da Cruz Neto, de Lima, Oliveira '98
Bento, Ferreira, Melo '17

1. Explicit: $x_{n+1} = \exp_{x_n}\big(-\tau\nabla f(x_n)\big)$

$R \geq 0$: (smoothness and) $\nabla^2 f \geq 0$ gives $O(1/n)$ convergence rates

$R \leq 0$: ? (nonlocal condition)

2. Implicit: $x_{n+1} = \arg\min_x f(x) + \frac{1}{2\tau} d^2(x, x_n)$

$R \leq 0$: $\nabla^2 f \geq 0$ gives $O(1/n)$ convergence rates

$R \geq 0$: if $\mathfrak{S}_c \geq 0$ then convexity of f on **Kim–McCann geodesics** gives $O(1/n)$ convergence rates

Wasserstein gradient flows, generalized geodesics (Ambrosio–Gigli–Savaré '05)

Thank you!