Gradient descent with a general cost

Flavien Léger

joint works with Pierre-Cyril Aubin-Frankowski

Outline

1. A new class of algorithms

2. Convergence theory

3. Applications

1. Gradient descent as minimizing movement

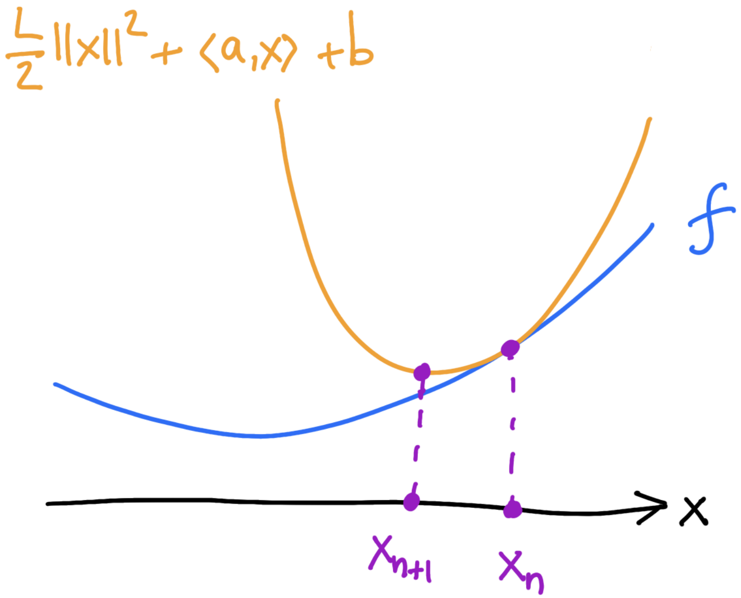
$$x_{n+1} = x_n - \frac{1}{L}\nabla f(x_n),$$

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

DEFINITION

f is L-smooth if

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



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

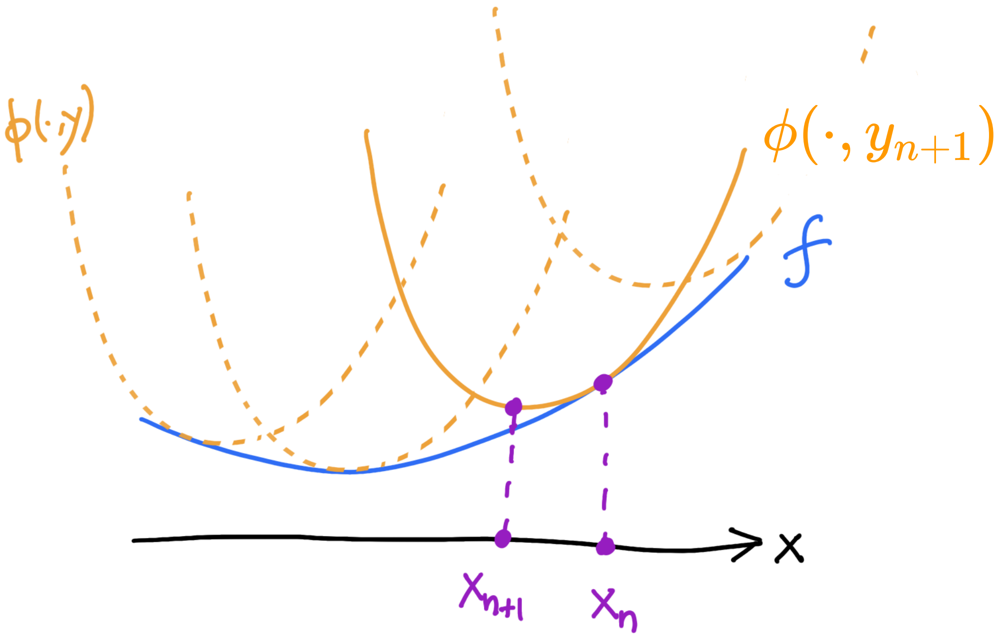
Two steps:

1) majorize: find the tangent parabola ("surrogate")

2) minimize: minimize the surrogate

The majorize step

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



Majorize step $\leftrightarrow$ y-update:

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

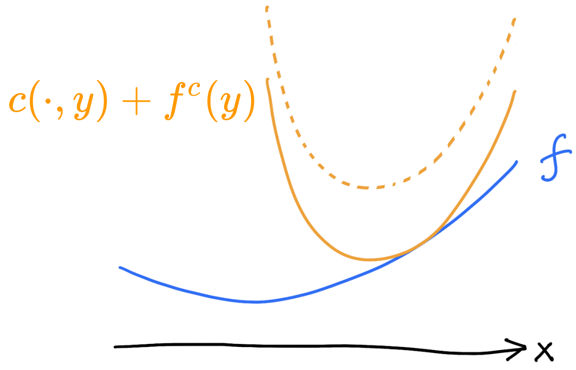
Minimize step $\leftrightarrow$ x-update:

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

General cost

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

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



DEFINITION ***c*-transform**

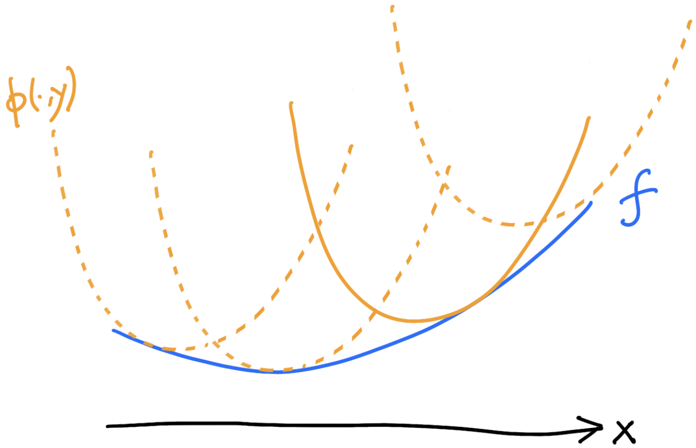
$$f^c(y) = \sup_{x \in X} c(x, y) - f(x)$$

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

DEFINITION

f is c-concave if

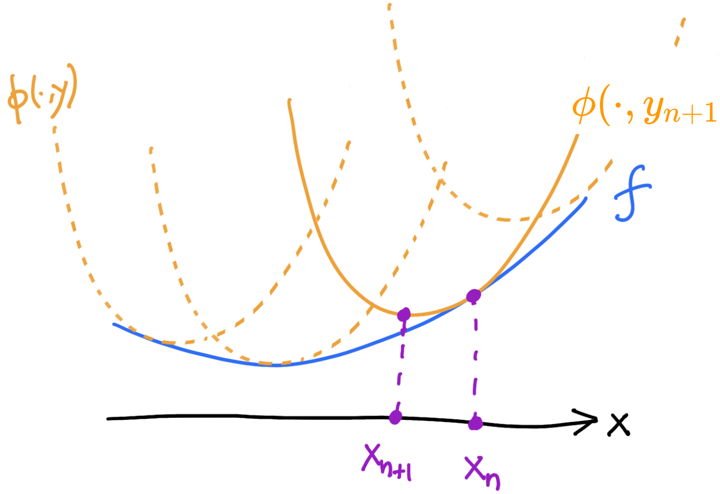
$$f(x) = \inf_{y \in Y} c(x, y) + f^c(y)$$



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"

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

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

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

Examples

- $c(x,y) = \overbrace{u(x) - u(y) - \langle \nabla u(y), x - y \rangle}^{=:u(x|y)}$: mirror descent

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

- $c(x,y) = u(y|x)$: natural gradient descent

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

Newton

- $c(x,y) = \frac{L}{2} d_M^2(x,y)$: Riemannian gradient descent

$-\nabla_x c(x,y) = \xi \Leftrightarrow y = \exp_x(\frac{1}{L}\xi)$

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

1. A new class of algorithms

2. Convergence theory

3. Applications

Cross-convexity and convergence rates

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

DEFINITION

f is λ-strongly c-cross-convex if

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

THEOREM (FL–PCAF '23)

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

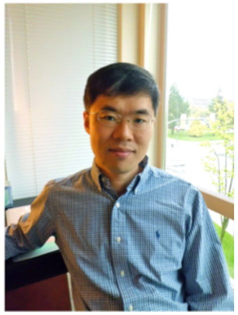
$$f(x_n) \leq f(x) + \frac{c(x, y_0) - c(x_0, y_0)}{n}.$$

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

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

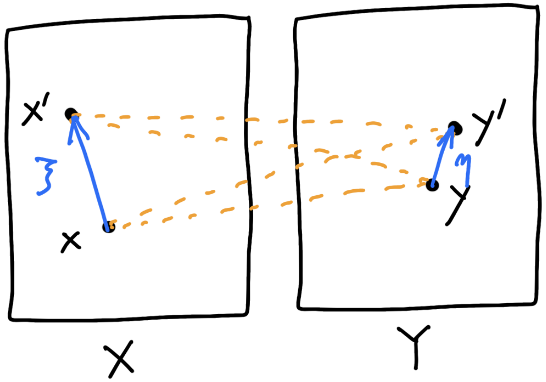
where $\Lambda := (1 - \lambda)^{-1} > 1$.

The Kim–McCann geometry



Kim and McCann ('10) introduced a pseudo-Riemannian metric on $X \times Y$

$$g_{(x,y)} = \begin{pmatrix} 0 & -\nabla^2_{xy} c(x,y) \\ -\nabla^2_{xy} c(x,y) & 0 \end{pmatrix}$$



$$[c(x,y') + c(x',y)] - [c(x,y) + c(x',y')]$$

➡ Kim–McCann geodesics

➡ Kim–McCann curvature: ***cross-curvature***, aka MTW tensor

A local criteria for cross-curvature

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 $\lambda > 0$. If $t \mapsto f(x(t)) - \lambda c(x(t), \bar{y})$ 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 class of algorithms

2. Convergence theory

3. Applications

Application: Newton's method

$c(x,y) = u(y|x) \longrightarrow$ NGD

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

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

Riemannian/metric space setting

$$\underset{x \in M}{\text{minimize}}\, f(x) \qquad\qquad c(x,y) = \tfrac{1}{2\tau} d^2(x,y)$$

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 nonneg. cross-curv, 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!