Generic Acceleration Schema Beyond Convexity

Courtney Paquette
Joint work with D. Drusvyatskiy (UW), H. Lin (Inria),
J. Mairal (Inria), and Z. Harchaoui (UW)

Department of Mathematics
University of Washington (Seattle)

INFORMS 2017
October 24, 2017

Generic acceleration of large finite sum problem

The **large finite sum problem**

$$\min_x \; f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x) + \psi(x)$$

Remarks

- Measures empirical risk in machine learning
- Fast incremental methods when f is convex
- When f is nonconvex remains largely open

Generic acceleration of large finite sum problem

The **large finite sum problem**

$$\min_{x} \; f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x) + \psi(x)$$

Remarks

- Measures empirical risk in machine learning
- Fast incremental methods when f is convex
- When f is nonconvex remains largely open

Two questions from nonconvex optimization

1. Apply a method, $\mathcal{M}$, for convex optimization to a nonconvex;
2. Design an convexity adapting algorithm which does not need to know whether the objective function is convex.

Background: proximal point methods

$\boxed{\textbf{Our goal:} \quad \min_x \; f(x), \quad f \text{ is } \mu\text{-strongly convex}}$

Proximal-point method (Martinet '70,72, Rockafellar '76):

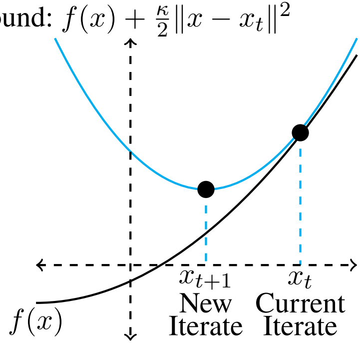
$$x_{t+1} = \operatorname*{argmin}_x \; f(x) + \frac{\kappa}{2}\|x - x_t\|^2$$

Background: proximal point methods

Our goal: $\min_x \ f(x)$, f is μ-strongly convex

Proximal-point method (Martinet '70,72, Rockafellar '76):

$$x_{t+1} = \operatorname*{argmin}_x \ f(x) + \frac{\kappa}{2}\|x - x_t\|^2$$



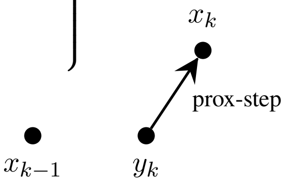
Complexity: $f(x_t) - f^* \sim \min\left\{\frac{\kappa}{t}, \left(1 - \frac{\mu}{\kappa}\right)^t\right\}$

Background: proximal point methods

Accelerated proximal-point method:
(Nesterov '83, Güler '92, Beck-Teboulle '09)

$$\left\{\begin{array}{l} \text{(Prox-step)} \quad x_t = \underset{x}{\operatorname{argmin}} \left\{ f(x) + \frac{\kappa}{2}\|x - y_{t-1}\|^2 \right\} \\ \text{Solve } \alpha_t^2 = (1-\alpha_t)\alpha_{t-1}^2 + \dfrac{\mu}{\mu+\kappa}\alpha_t \\ y_t = x_t + \beta_k(x_t - x_{t-1}) \\ \beta_t = \dfrac{\alpha_{t-1}(1-\alpha_{t-1})}{\alpha_{t-1}^2 + \alpha_t} \end{array}\right\}$$

Complexity:
$f(x_t) - f^* \sim \min\left\{\frac{\kappa}{t^2}, \left(1 - \sqrt{\frac{\mu}{\kappa}}\right)^t\right\}$

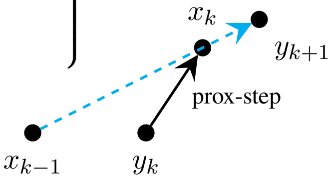


Background: proximal point methods

Accelerated proximal-point method:
(Nesterov '83, Güler '92, Beck-Teboulle '09)

$$
\left\{
\begin{array}{c}
\text{(Prox-step)} \quad x_t = \operatorname*{argmin}_x \left\{ f(x) + \frac{\kappa}{2} \|x - y_{t-1}\|^2 \right\} \\
\text{Solve } \alpha_t^2 = (1 - \alpha_t)\alpha_{t-1}^2 + \dfrac{\mu}{\mu + \kappa}\alpha_t \\
y_t = x_t + \beta_k (x_t - x_{t-1}) \\
\beta_t = \dfrac{\alpha_{t-1}(1 - \alpha_{t-1})}{\alpha_{t-1}^2 + \alpha_t}
\end{array}
\right\}
$$

Complexity:
$f(x_t) - f^* \sim \min \left\{ \frac{\kappa}{t^2}, \left(1 - \sqrt{\frac{\mu}{\kappa}}\right)^t \right\}$

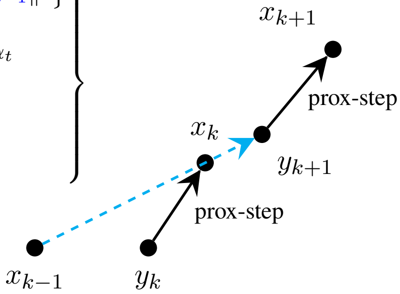


Background: proximal point methods

Accelerated proximal-point method:
(Nesterov '83, Güler '92, Beck-Teboulle '09)

$$\left\{\begin{array}{l} \text{(Prox-step)} \quad x_t = \underset{x}{\operatorname{argmin}} \left\{ f(x) + \frac{\kappa}{2}\|x - y_{t-1}\|^2 \right\} \\ \text{Solve } \alpha_t^2 = (1-\alpha_t)\alpha_{t-1}^2 + \dfrac{\mu}{\mu+\kappa}\alpha_t \\ y_t = x_t + \beta_k(x_t - x_{t-1}) \\ \beta_t = \dfrac{\alpha_{t-1}(1-\alpha_{t-1})}{\alpha_{t-1}^2 + \alpha_t} \end{array}\right\}$$

Complexity:
$f(x_t) - f^* \sim \min\left\{ \frac{\kappa}{t^2}, \left(1 - \sqrt{\frac{\mu}{\kappa}}\right)^t \right\}$

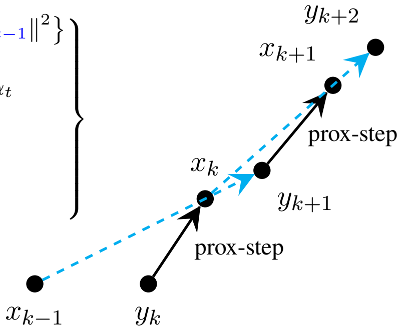


Background: proximal point methods

Accelerated proximal-point method:
(Nesterov '83, Güler '92, Beck-Teboulle '09)

$$\left\{\begin{array}{c} \text{(Prox-step)} \quad x_t = \underset{x}{\operatorname{argmin}} \left\{ f(x) + \frac{\kappa}{2}\|x - y_{t-1}\|^2 \right\} \\ \text{Solve } \alpha_t^2 = (1-\alpha_t)\alpha_{t-1}^2 + \frac{\mu}{\mu+\kappa}\alpha_t \\ y_t = x_t + \beta_k(x_t - x_{t-1}) \\ \beta_t = \frac{\alpha_{t-1}(1-\alpha_{t-1})}{\alpha_{t-1}^2 + \alpha_t} \end{array}\right\}$$

Complexity:
$f(x_t) - f^* \sim \min\left\{\frac{\kappa}{t^2}, \left(1 - \sqrt{\frac{\mu}{\kappa}}\right)^t\right\}$



Incremental methods

The problem

$$\min_x \ f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) + \psi(x),$$

- f μ-strongly convex and has bounded-level sets.
- $f_i \colon \mathbb{R}^p \to \mathbb{R}$ is L-smooth.
- $\psi \colon \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ may nonsmooth.

Incremental methods

The problem

$$\min_x \; f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) + \psi(x),$$

- f μ-strongly convex and has bounded-level sets.
- $f_i \colon \mathbb{R}^p \to \mathbb{R}$ is L-smooth.
- $\psi \colon \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ may nonsmooth.

	Progress	Efficiency
Incremental methods	$\mathbb{E}[f(x_t)] - f^* \le \varepsilon$	$\left(n + \frac{L}{\mu}\right) \cdot \ln \frac{1}{\varepsilon}$

SAG (Schmidt, Le Roux-Bach '13), SVRG (Johnson, Zhang '13), SAGA (Defazio, Bach, Lacoste-Julien '13), etc.

Convex Catalyst revisited: Accelerating methods

The problem

$$\min_x \ f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) + \psi(x),$$

- f convex
- $f_i : \mathbb{R}^p \to \mathbb{R}$ is smooth.
- $\psi : \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ may be nonsmooth.

Question: Given a non-optimal incremental method $\mathcal{M}$, can we make it optimal?

Convex Catalyst revisited: Accelerating methods

The problem

$$\min_x \; f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) + \psi(x),$$

- f convex
- $f_i \colon \mathbb{R}^p \to \mathbb{R}$ is smooth.
- $\psi \colon \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ may be nonsmooth.

Question: Given a non-optimal incremental method $\mathcal{M}$, can we make it optimal? Yes!

Idea: Wrap the method $\mathcal{M}$ within Nesterov's accelerated scheme.

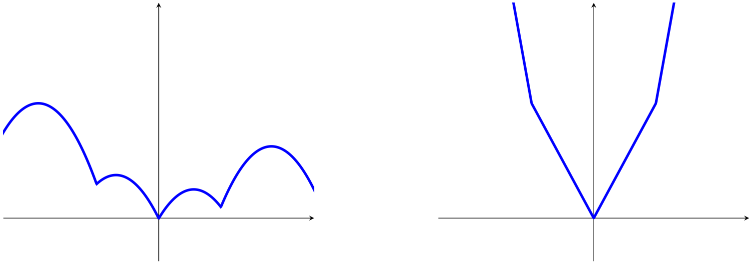
Main result (Lin, Mairal, Harchaoui, '15)

Other approaches included (Frostig et al. '15), (Shalev-Schwartz, Zhang '14), (Lan, Zhou, '15), (Allen-Zhu '16)

Tools for nonconvex and nonsmooth optimization

Definition I: A function $f : \mathbb{R}^p \to \overline{\mathbb{R}}$ is ***ρ*-weakly convex** if

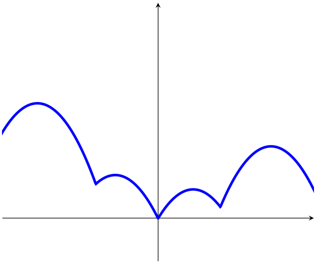
$$x \mapsto f(x) + \frac{\rho}{2}\|x\|^2 \quad \text{is convex.}$$

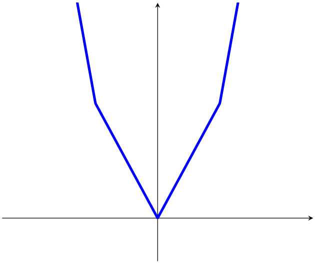


Tools for nonconvex and nonsmooth optimization

Definition I: A function $f : \mathbb{R}^p \to \overline{\mathbb{R}}$ is ***ρ*-weakly convex** if

$$x \mapsto f(x) + \frac{\rho}{2}\|x\|^2 \quad \text{is convex.}$$





Definition II: **subdifferential** of f is

$$\partial f(x) := \left\{ v \in \mathbb{R}^p \ : \ f(y) \geq f(x) + v^T(y - x) + o(\|y - x\|) \quad \forall y \in \mathbb{R}^p \right\}.$$

Seek points where

$$\mathrm{dist}(0, \partial f(x)) \leq \varepsilon.$$

Nonconvex catalyst: design a generic scheme

Our problem:

$$\min_{x \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} f_i(x) + \psi(x)$$

- $f_i : \mathbb{R}^p \to \mathbb{R}$ is smooth, $\boldsymbol{L}$-Lipschitz continuous gradient
- $\psi : \mathbb{R}^p \to \overline{\mathbb{R}}$ may be nonsmooth
- f is $\boldsymbol{\rho}$-weakly convex

Nonconvex catalyst: design a generic scheme

Our problem:

$$\min_{x \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} f_i(x) + \psi(x)$$

- $f_i : \mathbb{R}^p \to \mathbb{R}$ is smooth, $\boldsymbol{L}$-Lipschitz continuous gradient
- $\psi : \mathbb{R}^p \to \overline{\mathbb{R}}$ may be nonsmooth
- f is $\boldsymbol{\rho}$-weakly convex

Examples

1. Robust penalties (e.g. MCP, SCAD)
2. Neural networks

Our Problem

Two questions from nonconvex optimization

1. Apply a method, $\mathcal{M}$, for convex optimization to a nonconvex;
2. Design an convexity adapting algorithm which does not need to know whether the objective function is convex.

4WD-Catalyst : main idea

Build subproblems of the form:

$$\min_x \; f_\kappa(x; y) := f(x) + \frac{\boldsymbol{\kappa}}{2} \|x - y\|^2 .$$

$\boldsymbol{\kappa}$ is called the smoothing parameter

4WD-Catalyst : main idea

Build subproblems of the form:

$$\min_x \; f_\kappa(x;y) := f(x) + \frac{\boldsymbol{\kappa}}{2}\|x-y\|^2 .$$

$\boldsymbol{\kappa}$ is called the smoothing parameter

Two Steps

Step 1: Accelerated proximal point. Using $\mathcal{M}$ solve

$$y_k = \alpha_k v_{k-1} + (1-\alpha_k)x_{k-1}$$
$$\tilde{x}_k \approx \operatorname*{argmin}_x f_\kappa(x;y_k)$$
$$v_k = x_{k-1} + \frac{1}{\alpha_k}(\tilde{x}_k - x_{k-1})$$

where $\alpha_{k+1} \in (0,1)$ coefficients satisfying $(1-\alpha_{k+1})/\alpha_{k+1}^2 = 1/\alpha_k^2$.

Step 2: Proximal point. Using $\mathcal{M}$ solve

$$\bar{x}_k \approx \operatorname*{argmin}_x f_\kappa(x;x_{k-1})$$

Picking the best. Choose x_k such that

$$f(x_k) \leq \min\{f(\bar{x}_k), f(\tilde{x}_k)\} .$$

Similar to (Ghadimi-Lan '15)

Linear conv. of $\mathcal{M}$ for strongly-convex problems

Assumptions: For any $\kappa > \rho$, $\exists A_\kappa \geq 0, \tau_\kappa \in (0, 1)$:

1. $\forall z_0 \in \mathbb{R}^p$, the iters. $\{z_t\}_{t\geq 1}$ solve by $\mathcal{M}$:

$$\operatorname{dist}^2(0, \partial f_\kappa(z_t; y)) \leq A_\kappa(1 - \boldsymbol{\tau_\kappa})^t \big(f_\kappa(z_0; y) - f^*_\kappa(y)\big). \quad (1)$$

2. τ_κ, $A_\kappa \uparrow$ in κ.

Remarks

- (1) holds in $\mathbb{E}$;
- Gradient descent, SAGA, SVRG, satisfy these.

If $\boldsymbol{\kappa > \rho}$, then $\quad f_\kappa(x; y) := f(x) + \frac{\kappa}{2} \|x - y\|^2$

is $(\boldsymbol{\kappa - \rho})$-strongly convex!

Adaptive stopping criteria

What do we mean by

$$z^+ \approx \operatorname*{argmin}_x f_\kappa(x;y) := f(x) + \frac{\kappa}{2}\|x-y\|^2 \quad ?$$

Stopping criteria for the subproblems:

1. Descent condition: $f_\kappa(z^+;y) \leq f_\kappa(y;y)$;
2. Adaptive stationary condition:

$$\operatorname{dist}\big(0, \partial f_\kappa(z^+;y)\big) \leq \boldsymbol{\kappa} \left\|z^+ - y\right\|.$$

Adaptive stopping criteria

What do we mean by

$$z^+ \approx \operatorname*{argmin}_x f_\kappa(x;y) := f(x) + \frac{\kappa}{2}\|x-y\|^2 \quad ?$$

Stopping criteria for the subproblems:

1. Descent condition: $f_\kappa(z^+;y) \le f_\kappa(y;y)$;
2. Adaptive stationary condition:

$$\operatorname{dist}\big(0, \partial f_\kappa(z^+;y)\big) \le \boldsymbol{\kappa} \left\|z^+ - y\right\|.$$

For accelerated proximal point, we use

$$\operatorname{dist}\big(0, \partial f_\kappa(z^+;y)\big) \le \frac{\boldsymbol{\kappa}}{\boldsymbol{k+1}} \left\|z^+ - y\right\|.$$

Difficulty: In the nonconvex setting, $\mathcal{M}$ may "never" terminate!

*Exploit warm starts!

4WD-Catalyst

Step $k \geq 1$

1. **Proximal Step.** Compute $\bar{x}_k$ using $\mathcal{M}$ for ***T* iter.** until

$$\bar{x}_k \approx \operatorname*{argmin}_x f_\kappa(x; x_{k-1}) \quad \text{where}$$

$$\text{dist}(0, \partial f_\kappa(\bar{x}_k; x_{k-1})) < \kappa \|\bar{x}_k - x_{k-1}\| \;\&\; f_\kappa(\bar{x}_k; x_{k-1}) \leq f_\kappa(x_{k-1}; x_{k-1}).$$

2. **Accelerated Step.** Update y_k and compute $\tilde{x}_k$ using $\mathcal{M}$ for ***S* iter.** until:

$$\tilde{x}_k \approx \operatorname*{argmin}_x f_\kappa(x; y_k) \quad \text{where} \quad \text{dist}(0, \partial f_\kappa(\tilde{x}_k; y_k)) < \frac{\kappa}{k+1} \|\tilde{x}_k - y_k\|.$$

3. Update v_k and α_{k+1}.
4. **Picking the best.** Choose x_k:

$$f(x_k) \leq \min\{f(\bar{x}_k), f(\tilde{x}_k)\}.$$

Applications

	Nonconvex		Convex	
	Original	**4WD-Catalyst**	Original	**4WD-Catalyst**
Grad. Des.	$O\left(n\frac{L}{\varepsilon^2}\right)$	$\tilde{O}\left(n\frac{L}{\varepsilon^2}\right)$	$O\left(n\frac{L}{\varepsilon}\right)$	$\tilde{O}\left(n\sqrt{\frac{L}{\varepsilon}}\right)$
SVRG*	not avail.	$\tilde{O}\left(n\frac{L}{\varepsilon^2}\right)$	not avail.	$\tilde{O}\left(\sqrt{n}\sqrt{\frac{L}{\varepsilon}}\right)$
SAGA**	not avail.	$\tilde{O}\left(n\frac{L}{\varepsilon^2}\right)$	$O\left(n\frac{L}{\varepsilon}\right)$	$\tilde{O}\left(\sqrt{n}\sqrt{\frac{L}{\varepsilon}}\right)$
Coord. Des.	not avail.	$\tilde{O}\left(p^2\frac{L_{\max}}{\varepsilon^2}\right)$	$\mathcal{O}\left(p\frac{L_{\max}}{\varepsilon}\right)$	$\tilde{O}\left(p\sqrt{\frac{L_{\max}}{\varepsilon}}\right)$

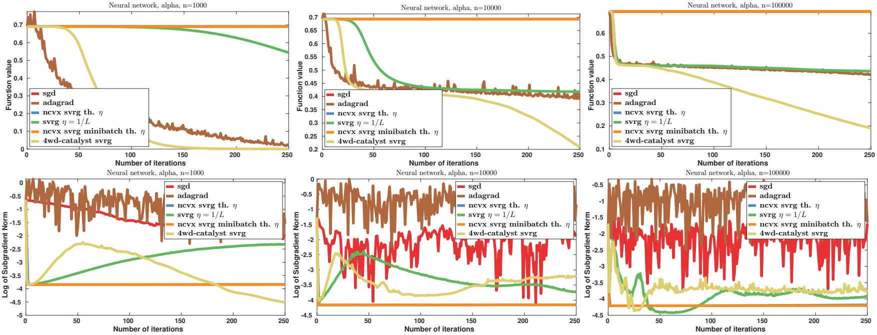
- Nonconvex–convergence stated in $\text{dist}(0, \partial f(x)) < \varepsilon$;
- Convex–convergence stated in $f(x) - f^* < \varepsilon$.

*SVRG (Johnson, Zhang '13), **SAGA (Defazio, Bach, Lacoste-Julien '13)

Experiments: neural networks

Given data $\{(a_i, b_i)\}_{i=1}^n$

$$\min_{W_1 \in \mathbb{R}^{n\times d}, W_2 \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \log\left(1 + \exp\left(-b_i\left(W_2^T \sigma(W_1^T a_i)\right)\right)\right)$$



Thank you!