Proximal Method with Contractions for Smooth Convex Optimization

Nikita Doikov Yurii Nesterov

Catholic University of Louvain, Belgium

Grenoble
September 23, 2019

Plan of the Talk

1. Proximal Method with Contractions

2. Application to Second-Order Methods

3. Numerical Example

Plan of the Talk

1. Proximal Method with Contractions

2. Application to Second-Order Methods

3. Numerical Example

$$f^* = \min_{x \in \mathbb{R}^n} f(x)$$

Proximal Method:

$$x_{k+1} = \operatorname*{argmin}_{y \in \mathbb{R}^n} \left\{ f(y) + \frac{1}{2a_{k+1}} \|y - x_k\|^2 \right\}.$$

[Rockafellar, 1976]

- If f is convex, the objective of the subproblem $h_{k+1}(y) = f(y) + \frac{1}{2a_{k+1}} \|y - x_k\|^2$ is strongly convex.
- Let f has Lipschitz gradient with constant L_1. Gradient Method needs $\tilde{O}\big(a_{k+1} L_1\big)$ iterations to minimize h_{k+1}.
- It is enough to use for x_{k+1} an inexact minimizer of h_{k+1}.

[Solodov-Svaiter, 2001; Schmidt-Roux-Bach, 2011; Salzo-Villa, 2012]

Set $a_{k+1} = \frac{1}{L_1}$. Then $f(\bar{x}_k) - f^* \leq \frac{L_1 \|x_0 - x^*\|^2}{2k}$.

Accelerated Proximal Method

Denote $A_k \stackrel{\mathrm{def}}{=} \sum_{i=1}^k a_i$. Two sequences: $\{x_k\}_{k\geq 0}$, and $\{v_k\}_{k\geq 0}$.

Initialization: $v_0 = x_0$.

Iterations, $k \geq 0$:

1. Put $y_{k+1} = \frac{a_{k+1}v_k + A_k x_k}{A_{k+1}}$.
2. Compute $x_{k+1} = \operatorname*{argmin}_{y\in\mathbb{R}^n}\Big\{ f(y) + \frac{A_{k+1}}{2a_{k+1}^2}\|y - y_{k+1}\|^2 \Big\}$.
3. Put $v_{k+1} = x_{k+1} + \frac{A_k}{a_{k+1}}(x_{k+1} - x_k)$.

Set $\frac{a_{k+1}^2}{A_{k+1}} = \frac{1}{L_1}$. Then

$$f(x_k) - f^* \;\leq\; \frac{8L_1\|x_0 - x^*\|^2}{3(k+1)^2}.$$

[Nesterov, 1983; Güler, 1992; Lin-Mairal-Harchaoui, 2015]

- *A Universal Catalyst for First-Order Optimization.*
- What about **Second-Order** Optimization?

Iterations, $k \geq 0$:

1. Compute $v_{k+1} = \operatorname*{argmin}\limits_{y \in \mathbb{R}^n} \left\{ A_{k+1} f\left(\frac{a_{k+1} y + A_k x_k}{A_{k+1}}\right) + \beta_d(v_k; y) \right\}$.

2. Put $x_{k+1} = \frac{a_{k+1} v_{k+1} + A_k x_k}{A_{k+1}}$.

$\beta_d(x; y)$ is Bregman Divergence.

Basic setup: $\beta_d(x; y) = \frac{1}{2}\|y - x\|^2$. Then

$$A_{k+1} f\left(\frac{a_{k+1} y + A_k x_k}{A_{k+1}}\right) + \frac{1}{2}\|y - v_k\|^2 \;=\; A_{k+1}\left(f(\tilde{y}) + \frac{A_{k+1}}{2a_{k+1}^2}\|\tilde{y} - y_{k+1}\|^2 \right),$$

where $\tilde{y} \equiv \frac{a_{k+1} y + A_k x_k}{A_{k+1}}$ and $y_{k+1} \equiv \frac{a_{k+1} v_k + A_k x_k}{A_{k+1}}$.

- The same iteration as in *Accelerated Proximal Method*.
- Generalization to arbitrary prox-function $d(\cdot)$.

Let $d(y)$ be a convex differentiable function. Denote **Bregman Divergence** of $d(\cdot)$, centered at x as

$$\beta_d(x;y) \stackrel{\mathrm{def}}{=} d(y) - d(x) - \langle \nabla d(x), y - x \rangle \geq 0.$$

- Mirror Descent [Nemirovski-Yudin, 1979]
- Gradient Methods with Relative Smoothness [Lu-Freund-Nesterov, 2016; Bauschke-Bolte-Teboulle, 2016]

Consider regularization of convex $g(\cdot)$ by Bregman Divergence:

$$h(y) \equiv g(y) + \beta_d(v;y).$$

Main Lemma. $T = \operatorname*{argmin}_{y \in \mathbb{R}^n} h(y)$. Then

$$h(y) \geq h(T) + \beta_d(T;y).$$

Proximal Method with Contractions: the Main Idea

We want, for all $y \in \mathbb{R}^n$:

$$\beta_d(x_0; y) + A_k f(y) \quad \geq \quad \beta_d(v_k; y) + A_k f(x_k). \qquad \textbf{(\$)}$$

How to propagate it to $k+1$? Denote $a_{k+1} \stackrel{\mathrm{def}}{=} A_{k+1} - A_k > 0$.

$$\begin{aligned}
\beta_d(x_0; y) + A_{k+1} f(y) \quad &\equiv \quad \beta_d(x_0; y) + A_k f(y) + a_{k+1} f(y) \\
&\stackrel{(\$)}{\geq} \quad \beta_d(v_k; y) + A_k f(x_k) + a_{k+1} f(y) \\
&\geq \quad \beta_d(v_k; y) + A_{k+1} f\left(\tfrac{a_{k+1} y + A_k x_k}{A_{k+1}}\right) \equiv h_{k+1}(y).
\end{aligned}$$

Let $\boldsymbol{v_{k+1} = \operatorname*{argmin}_{y \in \mathbb{R}^n} h_{k+1}(y)}$. Then, by the Main Lemma,

$$\begin{aligned}
h_{k+1}(y) \quad &\geq \quad h_{k+1}(v_{k+1}) + \beta_d(v_{k+1}; y) \\
&\geq \quad A_{k+1} f\underbrace{\left(\frac{a_{k+1} v_{k+1} + A_k x_k}{A_{k+1}}\right)}_{\equiv x_{k+1}} + \beta_d(v_{k+1}; y).
\end{aligned}$$

Iterations, $k \geq 0$:

1. Compute $v_{k+1} = \underset{y \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ A_{k+1} f\left(\frac{a_{k+1} y + A_k x_k}{A_{k+1}}\right) + \beta_d(v_k; y) \right\}$.

2. Put $x_{k+1} = \frac{a_{k+1} v_{k+1} + A_k x_k}{A_{k+1}}$.

Rate of convergence:

$$f(x_k) - f^* \quad \leq \quad \frac{\beta_d(x_0; x^*)}{A_k}.$$

Questions:

- How to choose A_k? Prox-function $d(\cdot)$?
- How to compute v_{k+1}?

Plan of the Talk

1. Proximal Method with Contractions

2. Application to Second-Order Methods

3. Numerical Example

$$h^* = \min_{x \in \mathbb{R}^n} h(x)$$

h is convex, with Lipschitz continuous Hessian:

$$\|\nabla^2 h(x) - \nabla^2 h(y)\| \leq L_2 \|x - y\|.$$

Model of the objective

$$\Omega_M(x; y) \stackrel{\mathrm{def}}{=} h(x) + \langle \nabla h(x), y - x \rangle + \tfrac{1}{2} \langle \nabla^2 h(x)(y - x), y - x \rangle + \tfrac{M}{6} \|y - x\|^3$$

Iterations:

$$\boxed{z_{t+1} := \operatorname*{argmin}_{y \in \mathbb{R}^n} \Omega_M(z_t; y), \quad t \geq 0.}$$

Newton method with Cubic regularization [Nesterov-Polyak, 2006]

- Global convergence

$$h(z_t) - h^* \leq O\left(\tfrac{L_2 R^3}{t^2}\right).$$

Apply **Cubic Newton** to compute the Proximal Step:

$$h_{k+1}(y) \;\; \equiv \;\; A_{k+1} f\left(\frac{a_{k+1}y + A_k x_k}{A_{k+1}}\right) + \beta_d(v_k; y) \;\; \to \;\; \min_{y \in \mathbb{R}^n}$$

- Pick $d(x) = \frac{1}{3}\|x - x_0\|^3$.
- Uniformly convex objective: $\beta_h(x; y) \geq \frac{1}{6}\|y - x\|^3$. Linear rate of convergence for Cubic Newton:

$$h(z_t) - h^* \;\; \leq \;\; O\left(\exp\left(-\frac{t}{\sqrt{L_2}}\right)(h(z_0) - h^*)\right).$$

- Let v_{k+1} be inexact Proximal Step: $\|\nabla h_{k+1}(v_{k+1})\|_* \leq \delta_{k+1}$.

Theorem

$$f(x_k) - f^* \;\leq\; \frac{\left(3^{-2/3}\|x_0 - x^*\|^2 + 6^{1/3}\sum_{i=1}^{k}\delta_i\right)^{3/2}}{A_k}$$

- $O\left(\sqrt{L_2(h_{k+1})}\log\frac{1}{\delta_{k+1}}\right)$ iterations of Cubic Newton for step k.

Contracted objective: $g_{k+1}(y) \;\equiv\; A_{k+1} f\left(\frac{a_{k+1}y + A_k x_k}{A_{k+1}}\right)$.

Derivatives

1. $Dg_{k+1}(y) = a_{k+1} Df\left(\frac{a_{k+1}y + A_k x_k}{A_{k+1}}\right)$,
2. $D^2 g_{k+1}(y) = \frac{a_{k+1}^2}{A_{k+1}} D^2 f\left(\frac{a_{k+1}y + A_k x_k}{A_{k+1}}\right)$,
3. $D^3 g_{k+1}(y) = \frac{a_{k+1}^3}{A_{k+1}^2} D^3 f\left(\frac{a_{k+1}y + A_k x_k}{A_{k+1}}\right)$,

   $\ldots$

Notice: $D^{p+1} f \;\preceq L_p(f) \;\Rightarrow\; D^{p+1} g_{k+1} \;\preceq\; \frac{a_{k+1}^{p+1}}{A_{k+1}^p} L_p(f)$. Therefore,

if we have $\boxed{\frac{a_{k+1}^{p+1}}{A_{k+1}^p} \leq \frac{1}{L_p(f)}}$ then $L_p(g_{k+1}) \leq 1$.

- For Cubic Newton ($p = 2$) set $A_k = \frac{k^3}{L_2(f)}$. We obtain accelerated rate of convergence: $O\left(\frac{1}{k^3}\right)$.

Basic Method

$p = 1$: Gradient Method.

$p = 2$: Newton method with Cubic regularization.

$p = 3$: Third order methods (admits effective implementation) [Grapiglia-Nesterov, 2019].

$\ldots$

- Prox-function: $d(x) = \frac{1}{p+1}\|x - x_0\|^{p+1}$. Set $A_k = \frac{k^{p+1}}{L_p(f)}$.
- Let $\delta_k = \frac{c}{k^2}$.

Theorem

$$f(x_k) - f^* \quad \leq \quad O\left(\frac{L_p(f)\|x_0 - x^*\|^{p+1}}{k^{p+1}}\right).$$

- $O\left(\log \frac{1}{\delta_k}\right)$ steps of *Basic Method* every iteration.

Plan of the Talk

1. Proximal Method with Contractions

2. Application to Second-Order Methods

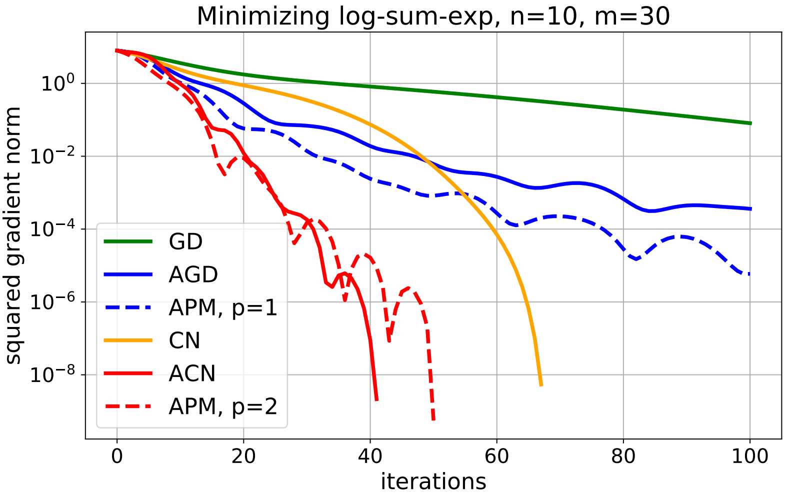
3. Numerical Example

$$\min_{x \in \mathbb{R}^n} f(x) \;=\; \log\left(\sum_{i=1}^{m} e^{\langle a_i, x\rangle}\right).$$

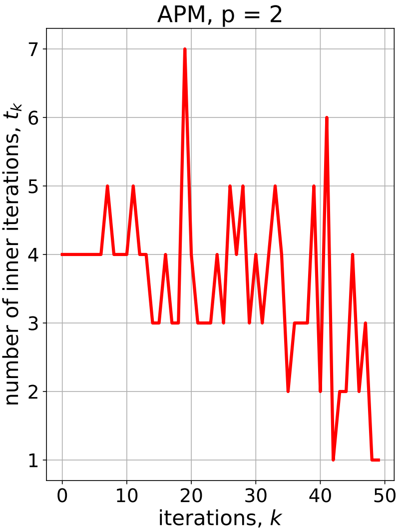
- $a_1, \dots, a_m \in \mathbb{R}^n$ — given data.
- Denote $B \equiv \sum\limits_{i=1}^{m} a_i a_i^T \succeq 0$, and use $\|x\| \equiv \langle Bx, x\rangle^{1/2}$.
- We have

$$L_1 \;\le\; 1, \qquad L_2 \;\le\; 2.$$

Log-sum-exp: convergence



Log-sum-exp: inner steps



Two ingredients

- Bregman divergence $\beta_d(v_k; y)$.
- Contraction operator

$$f(y) \quad \mapsto \quad f\left(\frac{a_{k+1}y + A_k x_k}{A_{k+1}}\right).$$

Direct acceleration vs. Proximal acceleration

- The rates are: $O\left(\frac{1}{k^{p+1}}\right)$ and $\tilde{O}\left(\frac{1}{k^{p+1}}\right)$, for the methods of order $p \geq 1$.
- In practice, the number of inner steps is a constant.
- Proximal acceleration is more general — useful for stochastic and distributed optimization.

Thank you for your attention!