Nearly optimal first-order method under Hölderian error bound: An adaptive proximal point approach

*Masaru Ito (Nihon University)

Mituhiro Fukuda (Tokyo Institute of Technology)

ICCOPT2019
August 6, 2019

Outline

1. Iteration complexity under the norm of the gradient mapping $\|g_L(x)\|$
2. Adaptive and nearly optimal first-order method for L-smooth functions
3. Hölderian error bound (HEB) condition
4. Adaptive and nearly optimal first-order method under HEB

Problem setting

Composite convex optimization problem

$$\text{minimize} \quad F(x) := f(x) + g(x) \quad \text{subject to} \quad x \in \mathbb{R}^n$$

- $f : \mathbb{R}^n \to \mathbb{R}$ is a L-smooth convex function:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall x, y$$

- $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a proper lower-semicontinuous convex function.
- F^*: the optimal value, X^*: the optimal solution set

- Proximal first-order method: An iterative method generates approximate solutions $\{x_k\}$ using $\nabla f(x)$ and $\text{prox}_g(x)$.
- $g(x)$ plays a role of a regularization term
  $\to$ Application to large-scale problems: data mining, machine learning, etc.

Iteration complexity

- Iteration Complexity = "Number of iterations to attain opt. measure $\leq \varepsilon$"

  Choice of optimality measure: $F(x) - F^*$, $\mathrm{dist}(x, X^*)$, etc.

Iteration complexity

- Iteration Complexity = "Number of iterations to attain opt. measure $\leq \varepsilon$"

  Choice of optimality measure: $F(x) - F^*$, $\mathrm{dist}(x, X^*)$, etc.

- Optimal iteration complexity is known for the measure $F(x) - F^*$:

  A family of accelerated gradient methods ensure the iteration complexity $\mathcal{O}\left(\sqrt{\frac{L\,\mathrm{dist}(x_0, X^*)^2}{\varepsilon}}\right)$ which is essentially unimprovable.

Iteration complexity

- Iteration Complexity = "Number of iterations to attain opt. measure $\leq \varepsilon$"

  Choice of optimality measure: $F(x) - F^*$, $\mathrm{dist}(x, X^*)$, etc.

- Optimal iteration complexity is known for the measure $F(x) - F^*$:

  A family of accelerated gradient methods ensure the iteration complexity $\mathcal{O}\left(\sqrt{\frac{L\,\mathrm{dist}(x_0, X^*)^2}{\varepsilon}}\right)$ which is essentially unimprovable.

- This work employs the measure: $\|g_L(x)\|$ $(= \|\nabla f(x)\|$ if $g \equiv 0)$

$$\text{Gradient mapping} \quad g_L(x) := L\left(x - \mathrm{prox}_{g/L}\left(x - \frac{1}{L}\nabla f(x)\right)\right),$$

$$\mathrm{prox}_{g/L}(y) := \operatorname*{argmin}_x \left(g(x) + \frac{L}{2}\|x - y\|^2\right).$$

  $g_L(x) = 0$ iff $x \in X^*$.

  $\|g_L(x)\|$ is available as a computable optimality measure.

Iteration complexity under gradient mapping norm

- A lower bound of the iteration complexity under the gradient norm is

$$\Omega\left(\sqrt{\frac{L\,\mathrm{dist}(x_0, X^*)}{\varepsilon}}\right)$$

for minimization $\min_x f(x)$ of L-smooth functions (Nemirovsky 1991).

Iteration complexity under gradient mapping norm

- A lower bound of the iteration complexity under the gradient norm is

$$\Omega\left(\sqrt{\frac{L\,\mathrm{dist}(x_0, X^*)}{\varepsilon}}\right)$$

  for minimization $\min_x f(x)$ of L-smooth functions (Nemirovsky 1991).
- Accelerated gradient methods can attain the iteration complexity $\mathcal{O}\left(\frac{\sqrt{L\,\mathrm{dist}(x_0, X^*)}}{\varepsilon^{2/3}}\right)$ (with a small modification).

Iteration complexity under gradient mapping norm

- A lower bound of the iteration complexity under the gradient norm is

$$\Omega\left(\sqrt{\frac{L\,\mathrm{dist}(x_0, X^*)}{\varepsilon}}\right)$$

for minimization $\min_x f(x)$ of L-smooth functions (Nemirovsky 1991).

- Accelerated gradient methods can attain the iteration complexity $\mathcal{O}\left(\frac{\sqrt{L\,\mathrm{dist}(x_0, X^*)}}{\varepsilon^{2/3}}\right)$ (with a small modification).
- A regularization technique (Nesterov 2012) attains near optimality

$$\mathcal{O}\left(\sqrt{\frac{L\,\mathrm{dist}(x_0, X^*)}{\varepsilon}}\log\frac{1}{\varepsilon}\right).$$

However, we require $\mathrm{dist}(x_0, X^*)$ to be known in advance.

Iteration complexity under gradient mapping norm

- A lower bound of the iteration complexity under the gradient norm is

$$\Omega\left(\sqrt{\frac{L\,\mathrm{dist}(x_0, X^*)}{\varepsilon}}\right)$$

for minimization $\min_x f(x)$ of L-smooth functions (Nemirovsky 1991).

- Accelerated gradient methods can attain the iteration complexity $\mathcal{O}\left(\frac{\sqrt{L\,\mathrm{dist}(x_0, X^*)}}{\varepsilon^{2/3}}\right)$ (with a small modification).

- A regularization technique (Nesterov 2012) attains near optimality

$$\mathcal{O}\left(\sqrt{\frac{L\,\mathrm{dist}(x_0, X^*)}{\varepsilon}}\log\frac{1}{\varepsilon}\right).$$

However, we require $\mathrm{dist}(x_0, X^*)$ to be known in advance.
$\rightarrow$ This requirement is reducible (This talk).

Regularization technique (Nesterov 2012)

Regularized problem:

$$\text{minimize}_x \quad F_{\sigma,x_0}(x) := F(x) + \frac{\sigma}{2}\|x - x_0\|^2, \qquad \sigma := \frac{\varepsilon}{2\,\mathrm{dist}(x_0, X^*)},$$

Optimal solution is $\mathrm{prox}_{F/\sigma}(x_0)$.
F_{σ,x_0} is σ-strongly convex for which we can apply accelerated gradient method:

$$F_{\sigma,x_0}(x_k) - \inf F_{\sigma,x_0} \le \mathcal{O}(1)L\|x_0 - \mathrm{prox}_{F/\sigma}(x_0)\|^2 \exp(-k\sqrt{\sigma/L})$$

Regularization scheme

Compute $\bar{x}$ ($\approx \mathrm{prox}_{F/\sigma}(x_0)$) via accelerated gradient method applied to F_{σ,x_0} and running $\mathcal{O}(1)\sqrt{L/\sigma}\log((L+\sigma)/\sigma)$ iterations.

We can show that

$$\|g_L(\bar{x})\| \le 2\sigma\,\mathrm{dist}(x_0, X^*) = \varepsilon.$$

Iteration complexity is

$$\mathcal{O}\left(\sqrt{\frac{L\,\mathrm{dist}(x_0, X^*)}{\varepsilon}} \log\frac{1}{\varepsilon}\right) : \quad \text{nearly optimal}$$

Adaptive regularization technique

Regularized problem:

$$\text{minimize}_x \quad F_{\sigma,x_0}(x) := F(x) + \frac{\sigma}{2}\|x - x_0\|^2, \qquad \sigma > 0. \quad \sigma := \frac{\varepsilon}{2\,\mathrm{dist}(x_0, X^*)}$$

Algorithm I: Adaptive regularization scheme

(a) Compute $\bar{x}$ ($\approx \mathrm{prox}_{F/\sigma}(x_0)$) via accelerated gradient method applied to F_{σ,x_0} and running $\mathcal{O}(1)\sqrt{L/\sigma}\log(L+\sigma)/\sigma$ iterations.
(b) If $\|g_L(\bar{x})\| > \varepsilon$ then $\sigma > \varepsilon/(2\,\mathrm{dist}(x_0, X^*))$ so restart (a) letting $\sigma \leftarrow \sigma/2$.
Otherwise, we obtain an ε-solution.

We can show that

$$\|g_L(\bar{x})\| \le 2\sigma\,\mathrm{dist}(x_0, X^*).$$

Main result I: Adaptive regularization with $\sigma := L$ ensures the iteration complexity

$$\mathcal{O}\left(\sqrt{\frac{L\,\mathrm{dist}(x_0, X^*)}{\varepsilon}}\log\frac{1}{\varepsilon}\right) : \quad \text{nearly optimal}$$

We do not require to know $\mathrm{dist}(x_0, X^*)$.

Hölderian error bound condition

Assumption: Hölderian Error Bound (HEB)

For the initial point $x_0 \in \mathbb{R}^n$, there exists $\kappa > 0$, $\rho \geq 1$ such that

$$F(x) - F^* \geq \kappa \, \mathrm{dist}(x, X^*)^{\rho}, \quad \forall x \text{ with } F(x) \leq F(x_0).$$

Hölderian error bound condition

Assumption: Hölderian Error Bound (HEB)

For the initial point $x_0 \in \mathbb{R}^n$, there exists $\kappa > 0$, $\rho \geq 1$ such that

$$F(x) - F^* \geq \kappa \,\mathrm{dist}(x, X^*)^{\rho}, \quad \forall x \text{ with } F(x) \leq F(x_0).$$

- Strong convexity implies HEB:

  F is μ-strongly convex $\iff F(x) \geq F(y) + F'(y; x-y) + \frac{\mu}{2}\|x-y\|^2, \ \forall x, y$

  $\Longrightarrow F(x) - F^* \geq \frac{\mu}{2}\mathrm{dist}(x, X^*)^2, \ \forall x$

  $\Longrightarrow$ HEB with $\kappa = \frac{\mu}{2}$, $\rho = 2$

Hölderian error bound condition

Assumption: Hölderian Error Bound (HEB)

For the initial point $x_0 \in \mathbb{R}^n$, there exists $\kappa > 0$, $\rho \geq 1$ such that

$$F(x) - F^* \geq \kappa\, \mathrm{dist}(x, X^*)^{\rho}, \quad \forall x \text{ with } F(x) \leq F(x_0).$$

- Strong convexity implies HEB:

  F is μ-strongly convex $\iff$ $F(x) \geq F(y) + F'(y; x-y) + \frac{\mu}{2}\|x-y\|^2, \ \forall x, y$

  $\implies F(x) - F^* \geq \frac{\mu}{2}\mathrm{dist}(x, X^*)^2, \ \forall x$

  $\implies$ HEB with $\kappa = \frac{\mu}{2}$, $\rho = 2$

- If F is a continuous convex, coercive, and semi-algebraic function, then, for any $x_0 \in \mathbb{R}^n$, there exists κ, ρ such that HEB holds.

  F is semi-algebraic $\iff$ graph (F) is semi-algebraic $\iff$
  graph $(F) = \bigcup_i^{\text{finite}} \bigcap_j^{\text{finite}} \{x : \mathrm{p}_{ij}(x) \leq 0\}$, $\quad p_{ij}$: a polynomial

Hölderian error bound condition

Assumption: Hölderian Error Bound (HEB)

For the initial point $x_0 \in \mathbb{R}^n$, there exists $\kappa > 0$, $\rho \geq 1$ such that

$$F(x) - F^* \geq \kappa \, \mathrm{dist}(x, X^*)^{\rho}, \quad \forall x \text{ with } F(x) \leq F(x_0).$$

Relation to Kurdyka-Łojasiewicz inequality (Bolte et al. 2017)

- Let f be a proper lower-semicontinuous convex function on X.
- Fix x_0 and $\rho \geq 1$.

Then, HEB holds for some $\kappa > 0$ if and only if there exists $c > 0$ such that

$$\mathrm{dist}(0, \partial f(x)) \geq c(f(x) - f^*)^{\alpha}, \quad \alpha = 1 - \frac{1}{\rho} \in [0, 1)$$

for all x with $f(x) \leq f(x_0)$.

Adaptive algorithms

	Problem class	Adaptive to	Measure
Nesterov '07 Lin & Xiao '15	μ-strong conv.	μ	$\|g_L(x)\|$
Fercoq & Qu '17	HEB with $\rho = 2$	κ	$\|g_L(x)\|$
Liu & Yang '17	HEB with known ρ	κ	$\|g_L(x)\|$
This work	HEB	κ, ρ	$\|g_L(x)\|$
Roulet & d'Aspremont '17 Renegar & Grimmer '18	HEB	κ, ρ	$F(x) - F^*$

Adaptive proximal-point strategy

Assumption: f is L-smooth (L is known) and admits HEB for some κ, ρ.

Algorithm II

$x_0 \in \mathbb{R}^n$, $\sigma > 0$ (regularization parameter). Set $x_0^+ := \mathrm{prox}_{g/L}(x_0 - \nabla f(x_0)/L)$

t-th stage $(t = 0, 1, 2, \ldots,)$:

(a) Compute $x_t^{(0)}, x_t^{(1)}, \ldots$ ($\approx \mathrm{prox}_{F/\sigma}(x_t^+)$) via accelerated gradient method applied to F_{σ, x_t^+}, starting from x_t^+, running K_t iterations, where

$$K_t := \mathcal{O}(1)\sqrt{L/\sigma}\log((L+\sigma)/\sigma), \quad F_{\sigma, x_t^+}(x) := F(x) + \frac{\sigma}{2}\|x - x_t^+\|^2$$

($*$) If $\|g_L(x_t^{(k)})\| \le \|g_L(x_t)\|/2$ holds at some k,

then set $x_{t+1} := x_t^{(k)}, \quad x_{t+1}^+ := \mathrm{prox}_{g/L}(x_{t+1} - \nabla f(x_{t+1})/L)$

and go to $(t+1)$-th stage.

(b) Set $\sigma \leftarrow \sigma/2$ and retry t-th stage.

Iteration complexity of the proposed method

Proposed method:

$x_{t+1} \leftarrow \text{Accelerated Gradient Method}\,(F_{\sigma, x_t^+},\ x_t^+,\ K_t) \approx \mathrm{prox}_{F/\sigma}(x_t^+)$

$x_{t+1}^+ := \mathrm{prox}_{g/L}\,(x_{t+1} - \nabla f(x_{t+1})/L)$

If $\|g_L(x_{t+1})\| \leq \|g_L(x_t)\|/2$ then, go to $(t+1)$-th stage.

Otherwise, set $\sigma \leftarrow \sigma/2$ and retry t-th stage.

Iteration complexity of the proposed method

Proposed method:

$x_{t+1} \leftarrow$ Accelerated Gradient Method $(F_{\sigma, x_t^+},\ x_t^+,\ K_t) \approx \mathrm{prox}_{F/\sigma}(x_t^+)$

$x_{t+1}^+ := \mathrm{prox}_{g/L}\left(x_{t+1} - \nabla f(x_{t+1})/L\right)$

If $\|g_L(x_{t+1})\| \leq \|g_L(x_t)\|/2$ then, go to $(t+1)$-th stage.

Otherwise, set $\sigma \leftarrow \sigma/2$ and retry t-th stage.

Main result II

Iteration complexity of the proposed method when σ is initialized by $\sigma := L$:

Case	$\rho = 1$	$1 < \rho < 2$	$\rho = 2$	$\rho > 2$
Convergence of $\|g_L(x_t)\|$	finite	superlinear	linear	sublinear
Complexity w.r.t. ε	const	$\mathcal{O}(\log\log\frac{1}{\varepsilon})$	$\mathcal{O}(\log\frac{1}{\varepsilon})^{*1}$	$\mathcal{O}(\varepsilon^{-\frac{\rho-2}{2(\rho-1)}}\log\frac{1}{\varepsilon})^{*2}$

$$(*1) = \mathcal{O}\left(\sqrt{\frac{L}{\kappa}}\log\frac{L}{\kappa}\log\frac{1}{\varepsilon}\right), \quad (*2) = \mathcal{O}\left(\sqrt{\frac{L}{\kappa^{\frac{1}{\rho-1}}\varepsilon^{\frac{\rho-2}{\rho-1}}}}\log\frac{1}{\varepsilon}\right)$$

Near optimality

Lemma

If F admits HEB for some $x_0 \in \mathbb{R}^n$, $\rho > 1$, $\kappa > 0$, then

$$F(x^+) - F^* \leq 2^{\frac{\rho}{\rho-1}} \left(\frac{1}{\kappa}\right)^{\frac{1}{\rho-1}} \|g_L(x)\|^{\frac{\rho}{\rho-1}}, \quad \forall x \text{ with } F(x) \leq F(x_0).$$

where $x^+ := \mathrm{prox}_{g/L}(x - \nabla f(x)/L)$.

- A lower complexity bound for $F(x) - F^*$ induces the one for $\|g_L(x)\|$.
- Lower iteration complexity bound for $F(x) - F^*$ is known in the case $g \equiv 0$ (In this case ρ must be ≥ 2) [Nemirovsky & Nesterov 1985].

Case	$\rho = 1$	$1 < \rho < 2$	$\rho = 2$	$\rho > 2$
Convergence of $\|g_L(x_t)\|$	finite	super-linear	linear	sublinear
Complexity w.r.t. ε	const	$\mathcal{O}(\log\log\frac{1}{\varepsilon})$	$\mathcal{O}(\log\frac{1}{\varepsilon})$	$\mathcal{O}(\varepsilon^{-\frac{\rho-2}{2(\rho-1)}}\log\frac{1}{\varepsilon})$
			nearly optimal	

- Proposed method also ensures the near optimality w.r.t. $F(x) - F^*$.

Summary

- Developed a simple adaptive proximal-point strategy of first-order method under the measure $\|g_L(x)\|$.
- For minimization L-smooth functions, we can achieve the iteration complesity $\mathcal{O}\left(\sqrt{\frac{L\,\mathrm{dist}(x_0,X^*)}{\varepsilon}}\log\frac{1}{\varepsilon}\right)$ without knowing $\mathrm{dist}(x_0, X^*)$.
- We can adapt to the HEB condition and attain nearly optimal complexity.

Future interest

- Nonsmooth case or weakly smooth case.
- Other error bound conditions.

References

- J. Bolte, T. P. Nguyen, J. Peypouquet, and B. W. Suter, From error bounds to the complexity of first-order descent methods for convex functions, *Math. Program.*, **165**, pp. 471–507, 2017.
- O. Fercoq and Z. Qu, Adaptive restart of accelerated gradient methods under local quadratic growth condition, arXiv:1709.02300, 2017.
- Qihang Lin and Lin Xiao, An adaptive accelerated proximal gradient method and its homotopy continuation for sparse optimization, *Comput. Optim. Appl.*, **60**, pp. 633–674, 2015.
- Mingrui Liu and Tianbao Yang, Adaptive accelerated gradient converging methods under Hölderian error bound condition, arXiv:1611.07609, 2017.
- A. S. Nemirovsky, On optimality of Krylov's information when solving linear operator equations, *Journal of Complexity*, **7**, pp. 121–130, 1991.
- A. Nemirovski and Y. Nesterov, Optimal methods of smooth convex optimization, *U.S.S.R. Comput. Maths. Math. Phys.*, **25**(2), pp. 21–30, 1985.

References

Y. Nesterov, Gradient methods for minimizing composite functions, *Mathematical Programming*, **140**, pp. 125–161, 2013.

J. Renegar and B. Grimmer, A Simple Nearly-Optimal Restart Scheme For Speeding-Up First Order Methods, arXiv:1803.00151, 2018.

V. Roulet and A. d'Aspremont, Sharpness, Restart and acceleration, in *Advances in Neural Information Processing Systems*, pp. 1119–1129, 2017.