First-Order Methods for Nonsmooth Convex Large-Scale Optimization, I: General Purpose Methods

Maziar Sanjabi

University of Minnesota

February 11, 2014

- A. Juditsky, and A. Nemirovski. *"First order methods for non smooth convex large-scale optimization, I: general purpose methods."* Optimization for Machine Learning (2010): 121-148.

Outline

First Order Methods

Outline

Outline

Outline

Outline

First Order Methods (FOM)

- The goal is to solve

$$f^* = \min_{x \in \mathcal{X}} \; f(x) \tag{1}$$

  within ϵ accuracy ($f(\hat{x}) - f^* < \epsilon$).
- What class of problems are we dealing with?
  - General convex $f \in \mathcal{F}$ (Probably non-smooth);
  - $\mathcal{X}$ closed convex (simple).
- First Order Oracle: A black box that takes x & gives you $f(x)$ and a $f'(x)$.
- FOM is an algorithm that given any $\epsilon > 0$
  - knows $\mathcal{F}, \mathcal{X}$;
  - does not know f, and only has access to oracle.

  After finite number of oracle calls should give $\hat{x} \in \mathcal{X}$, s.t. $f(\hat{x}) - f^* < \epsilon$.

Lower Complexity Bounds

Given $\mathcal{X}, \mathcal{F}, \epsilon$, what is the minimum number of Oracle calls an FOM needs to give an ϵ-accuracy solution.

Definitions:

- $\mathcal{X} \subset \mathbb{R}^n$;
- $\mathcal{B}_p(R) = \{x \in \mathbb{R}^n : \|x\|_p \leq R\}$
- $\mathcal{F}_p(L)$: set of convex Lipschitz function with given constant L.

Class	Complexity Bound	Achievable
$f \in \mathcal{F}_p(L)$, $\mathcal{X} \subset \mathcal{B}_p$, $p \in [1,2]$	$O(1)\min[n, L^2R^2/\epsilon^2]$	$O(1)(\ln(n))^{2/p-1}L^2R^2/\epsilon^2$
$f \in \mathcal{F}_\infty(L)$, $\mathcal{X} \subset \mathcal{B}_\infty$	$O(1)n\ln(LR/\epsilon)$	-
$f \in \mathcal{S}_2(L)$, $\mathcal{X} \subset \mathcal{B}_2$	$O(1)\min[n, \sqrt{LR^2/\epsilon}]$	$O(1)\sqrt{LR^2/\epsilon}$

FOM vs. higher order algorithms

FOM Cons:

- Not suitable for high accuracy.
- sub-linear convergence.
- Speed relies heavily on constant such as L and R.

FOM Pros:

- "Cheap" iteration.
- Almost dimension independent iteration complexity.
- Good for medium accuracy, large scale optimization.

Note that $L.R$ matters in convergence and depends on norm imposed.
No assumption on the structure of functions.

Mirror Descent (MD) Method

$$\min_{x \in \mathcal{X}} \; f(x) \tag{2}$$

- $\mathcal{X} \subset E = \mathbb{R}^n$ closed convex set.
- f convex Lipschitz (with respect to some norm).
- The problem is solvable.
- Conjugate norm imposed on linear functionals.

$$E^* : \; \|\xi\|_* = \max_x \{\langle \xi, x \rangle : \|x\| \leq 1\} \tag{3}$$

- Distance generating function $w(\cdot)$:

$$\langle w'(x) - w'(x'), x - x' \rangle \geq \|x - x'\|^2 \tag{4}$$

$V_x(u) = w(u) - w(x) - \langle w'(x), u - x \rangle$.

- $\Omega := \max_{x \in \mathcal{X}} w(x) - \min_{x \in \mathcal{X}} w(x) = \max_{u \in \mathcal{X}} V_{x_c}(u)$.

MD Method (Con'd)

Examples

- Choosing $w(x) = \frac{1}{2}\|x\|_F^2 \Rightarrow V_x(u) = \frac{1}{2}\|x - u\|_F^2$.
- When using $\|\cdot\|_1$ on space $\mathcal{X}$ that is probability simplex, $w(x) = \sum_{i=1}^n x_i \ln(x_i) \Rightarrow V_x(u) = \sum_{i=1}^N u_i \ln(\frac{u_i}{x_i})$.

Define

$$Prox_x(\xi) = \arg\min_{u \in \mathcal{X}} \{\langle \xi, u \rangle + V_x(u)\}. \tag{5}$$

MD algorithm

(a) Start with $x_1 = \arg\min_{x \in \mathcal{X}} w(x)$

(b) In each iteration set $x_{\tau+1} = Prox_{x_\tau}(\gamma_t f'(x_\tau))$, $\tau = 1, \cdots, t$

(c) Output $\bar{x}_t = [\sum_{i\tau=1}^t \gamma_\tau]^{-1} \sum_{\tau=1}^t \gamma_\tau x_\tau$ and $\bar{f}_t = f(\bar{x}_t)$.

Convergence of MD

Theorem 1

Suppose f is Lipschitz continuous on $\mathcal{X}$ with $L := \sup_{x\in\mathcal{X}} \|f'(x)\|_*$, then using MD we have

$$\bar{f}_t - f^* \le \frac{V_{x_1}(x_*) + \frac{L^2}{2}\sum_{\tau=1}^t \gamma_\tau^2}{\sum_{\tau=1}^t \gamma_\tau} \tag{6}$$

Remark 1

Choosing $\gamma_t = \gamma/[\|f'(x_t)\|_* \sqrt{t}]$, will give

$$\boxed{\bar{f}_t - f^* \le O(1)\,\left[\frac{V_{x_1}(x_*)}{\gamma} + \frac{\ln(t+1)\gamma}{2}\right] L t^{-1/2}}$$

Remark 2

Setting $\gamma_\tau = \frac{\sqrt{2\Omega}}{L\sqrt{t}}$ yields

$$\boxed{\bar{f}_t - f^* \le \frac{\sqrt{2\Omega}L}{\sqrt{t}}}$$

Proof

- Start with first order optimality condition of the update

$$\forall\, u \in \mathcal{X},\ \langle \gamma_\tau f'(x_\tau) - w'(x_\tau) + w'(x_{\tau+1}), u - x_{\tau+1} \rangle \geq 0 \tag{7}$$

- Massage it to get inequalities similar to

$$\begin{aligned} \gamma_\tau \langle f'(x_\tau), x_\tau - u \rangle \leq \\ V_{x_\tau}(u) - V_{x_{\tau+1}}(u) + [\gamma_\tau \langle f'(x_\tau), x_\tau - x_{\tau+1} \rangle - V_{x_\tau}(x_{\tau+1})] \end{aligned} \tag{8}$$

- Plug in x^* and use optimality of x^*, i.e. $f(x_\tau) - f^* \leq \langle f'(x_\tau), x_\tau - x^* \rangle$

$$\gamma_\tau (f(x_\tau) - f^*) \leq V_{x_\tau}(x^*) - V_{x_{\tau+1}}(x^*) + \delta_\tau \tag{9}$$

Proof (Con'd)

- Bounding δ_τ using strong convexity

$$\begin{aligned}\delta_\tau \leq & \gamma_\tau \langle f'(x_\tau), x_\tau - x_{\tau+1}\rangle - \frac{1}{2}\|x_\tau - x_{\tau+1}\|^2 \\ \leq & \gamma_\tau \|f'(x_\tau)\|_* \|x_\tau - x_{\tau+1}\| - \frac{1}{2}\|x_\tau - x_{\tau+1}\|^2 \leq \frac{\gamma_\tau^2}{2}\|f'(x_\tau)\|_*^2 \end{aligned} \quad (10)$$

- Combining results we get

$$\gamma_\tau (f(x_\tau) - f^*) \leq V_{x_\tau}(x^*) - V_{x_{\tau+1}}(x^*) + \frac{\gamma_\tau^2}{2} L^2 \quad (11)$$

Adding up these inequalities for $\tau = 1. \cdots, t$, and normalizing the result by $\sum_{\tau=1}^t \gamma_\tau$ + convexity of f

$$f(\bar{x}_t) - f^* \leq [\sum_{\tau=1}^t \gamma_\tau]^{-1} \sum_{\tau=1}^t \gamma_\tau f(x_\tau) \; - \; f^* \leq \frac{V_{x_1}(x_*) + \frac{L^2}{2}\sum_{\tau=1}^t \gamma_\tau^2}{\sum_{\tau=1}^t \gamma_\tau}$$

Mirror Descent with Stochastic Approximation

- We have a stochastic first order oracle.
- Each time you give it an x, it gives back $G(x, \xi)$; ξ *iid* for each call.
- $g(x) = \mathbb{E}_\xi\{G(x,\xi)\}$; sub-gradient estimation error $\|g(x) - f'(x)\|_* \le \mu$
- $\mathbb{E}\{\|G(x,\xi)\|_*^2\} \le L^2$.
- Same MD algorithm, replacing $f'(x_\tau)$ with $G(x_\tau, \xi_\tau)$.

Proposition 1

Using the Stochastic Mirror Descent Algorithm in t steps we get

$$\mathbb{E}\{f(\bar{x}_t) - f^*\} \le \frac{\Omega + \frac{L^2}{2}\sum_{\tau=1}^{t}\gamma_\tau^2}{\sum_{\tau=1}^{N}\gamma_\tau} + \mu D, \tag{12}$$

where $D = \max_{x,x' \in \mathcal{X}} \|x - x'\|$.

Proof

- Similar to proof of theorem 1,
$$\gamma_\tau \langle G(x_\tau, \xi_\tau), x_\tau - x_* \rangle \le V_{x_\tau}(x^*) - V_{x_{\tau+1}}(x^*) + \gamma_\tau^2 \frac{L^2}{2} \tag{13}$$
- Adding them up fro $\tau = 1, \cdots, t$, we get
$$\sum_{\tau=1}^{t} \gamma_\tau \langle G(x_\tau, \xi_\tau), x_\tau - x_* \rangle \le \Omega + \frac{L^2}{2} \sum_{\tau=1}^{t} \gamma_\tau^2. \tag{14}$$
- Taking expectation of LHS with respect to $\xi_1, \cdots, \xi_t$:
  - x_τ is a deterministic function of $\xi_1, \cdots, \xi_{\tau-1}$.
  - Given $\xi_1, \cdots, \xi_{\tau-1}$,

$$\mathbb{E}_{\xi_\tau}\{\langle G(x_\tau, \xi_\tau), x_\tau - x_* \rangle\} = \langle g(x_\tau), x_\tau - x_* \rangle \ge \langle f'(x_\tau), x_\tau - x_* \rangle - \mu D$$

Therefore,
$$\mathbb{E}\{f(\bar{x}_t) - f^*\} \le \frac{1}{\sum_{\tau=1}^{t} \gamma_\tau} \mathbb{E}\left\{ \sum_{\tau=1}^{t} \gamma_\tau \langle f'(x_\tau), x_\tau - x_* \rangle \right\}$$
$$\le \frac{\Omega + \frac{L^2}{2} \sum_{\tau=1}^{t} \gamma_\tau^2}{\sum_{\tau=1}^{t} \gamma_\tau} + \mu D \tag{15}$$

Other Extensions of MD

- MD could be modified in order to solve

$$\min_{x \in \mathcal{X}} f(x)$$
$$s.t.\ f_i(x) \leq 0,\ i = 1, \cdots, m. \tag{16}$$

with the same iteration complexity.

- MD could also be modified (in an algorithm with multiple restarts) to solve

$$\min_{x \in \mathcal{X}} f(x), \tag{17}$$

when f is strongly convex with convergence rate of $O(1/t)$.

- MD could also be modified to solve convex-concave saddle point problem

$$\inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \phi(x, y) \tag{18}$$

with similar convergence rate results.

Setting up MD for some examples

- $\|\cdot\|_2$: Then MD is equivalent to sub-gradient projection.
  - Projection to $\mathcal{X}$ might not be easy.
  - In some cases such as when $\mathcal{X} = \mathcal{B}_2$ or a box, then this projection is easy.
- $\|\cdot\|_1$: There are some choices for $w(\cdot)$:
  - When $\mathcal{X}$ is probability simplex, choosing w as Entropy function (Most common choice). MD equivalent to Multiplicative update.

$$\arg\min_{u\in\mathcal{X}} \langle \gamma g, u\rangle + \sum_{i=1}^{n} u_i \ln(u_i/x_i) \Rightarrow$$
$$u_i = \alpha e^{-\gamma g_i} x_i. \tag{19}$$

  - When $\mathcal{X} = \mathcal{B}_1$, $w(x) = 2e\ln(n)\sum_{i=1}^{n} |x_i|^{p(n)}$, $p(n) = 1 + \frac{1}{2\ln(n)}$
- For matrix case, the Schatten p-norm could be used. As discussed in the previous presentation.

Why geometry is important?

- Consider the case where we use MD with constant step size, then

$$\bar{f}_t - f^* \leq \frac{\sqrt{2\Omega}L}{\sqrt{t}} \tag{20}$$

- We focus on L and Ω.
- Assume two cases, when we use $\mathcal{B}_p$, $p = 1, 2$, then the relative efficiency of MD algorithms would be

$$\frac{\text{Eff(Eucl)}}{\text{Eff}(\ell_1)} = O(1).\frac{1}{n^{1-1/p}\sqrt{\ln(n)}}.\frac{\sup_{x\in\mathcal{X}} \|f'(x)\|_2}{\sup_{x\in\mathcal{X}} \|f'(x)\|_\infty} \tag{21}$$

- First one is in favor of ℓ_2-MD; Second one is in favor of ℓ_1-MD

$$1 \leq B = \frac{\sup_{x\in\mathcal{X}} \|f'(x)\|_2}{\sup_{x\in\mathcal{X}} \|f'(x)\|_\infty} \leq \sqrt{n},\ A = \frac{1}{n^{1-1/p}\sqrt{\ln(n)}} \leq 1 \tag{22}$$

- If $p = 2$, then $A.B \leq 1$, Euclidean MD will have better performance.
- If $p = 1$, then there is a good chance that $A.B \geq 1$, ℓ_1-MD will have better performance.

Thank You!