Mirror descent for constrained strongly convex optimization

A. Bayandina

Department of Control and Applied Mathematics
Moscow Institute of Physics and Technology

December 28, 2016

Problem Statement

$$f(x) \to \min_{x \in Q},$$
$$s.t.\;\; g(x) \leqslant 0.$$

- E is a n-dimensional real vector space
- $Q \subset E$ is a convex compact
- $f : Q \to \mathbb{R}$ and $g : Q \to \mathbb{R}$ are μ-strongly convex w.r.t. some norm $\|\cdot\|$ and subdifferentiable

$$f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|^2$$

Stochastic Setting

- $(\Omega, \mathcal{F}, P)$ is a probability space
- $\{\xi^k\}$ is a sequence of i.i.d random vectors, each ξ^k is $\mathcal{F}$-measurable

Stochastic Gradient Oracle

- $x^k \in Q \quad \mapsto \quad g(x^k), \quad \nabla_x f(x^k, \xi^k), \quad \nabla_x g(x^k, \xi^k)$
- $\mathbb{E}_{\xi^k}[\nabla_x f(x^k, \xi^k)] = \nabla f(x^k)$
- $\mathbb{E}_{\xi^k}[\nabla_x g(x^k, \xi^k)] = \nabla g(x^k)$

Proximal Setup

- $\|\cdot\|$ is a norm in E ($\|\cdot\|_*$ is a norm in E^*)
- Domain $Q \subset E$
- Distance-generating function $d(x) : Q \to \mathbb{R}$, continuous and 1-strongly convex w.r.t. $\|\cdot\|$
- $x^0 = \arg\min_{x \in Q} d(x)$
- d-radius of Q

$$\omega_n = \sup_{y \in Q} \frac{2V_{x^0}(y)}{\|y - x^0\|^2} \sim \ln(n)$$

Mirror Descent

- Bregman distance from $x \in Q_0$ to $y \in Q$

$$V_x(y) := d(y) - \langle \nabla d(x), y - x \rangle - d(x)$$

- Starting point $x^0 = \arg\min_{x \in Q} d(x)$
- 'Radius' of the set Q

$$\Theta^2 = \sup_{x,y \in Q} V_x(y)$$

- Proximal mapping operator

$$\mathrm{Mirr}_x(u) := \arg\min_{y \in Q} \{ V_x(y) + \langle u, y - x \rangle \}$$

Convex Case Algorithm

Algorithm 1 Mirror Descent

Require: h_f, h_g, ε_g

procedure $\textsc{Mirror}(x^0, N, \Theta^2)$
    initialize I as an empty set
    for $k \in \{1, \ldots, N\}$ **do**
        if $g(x^k) \leqslant \varepsilon_g$ **then**
            $x^{k+1} \leftarrow \mathrm{Mirr}_{x^k}(h_f \nabla_x f(x^k, \xi^k))$
            add k to I
        else
            $x^{k+1} \leftarrow \mathrm{Mirr}_{x^k}(h_g \nabla_x g(x^k, \xi^k))$
    return $\bar{x}^N = \frac{1}{|I|} \sum_{k \in I} x^k$

Convex Case. Probability of Large Deviations

Theorem 1

Suppose for all $x \in Q$ and $\xi \in \{\xi^k\}$ it holds that

$$\|\nabla_x f(x,\xi)\|_*^2 \leqslant M_f^2, \quad \|\nabla_x g(x,\xi)\|_*^2 \leqslant M_g^2.$$

Then, if set $h_g = \dfrac{\varepsilon_g}{M_g^2}$, $h_f = \dfrac{\varepsilon_g}{M_f M_g}$, $\varepsilon_f = \dfrac{M_f}{M_g}\varepsilon_g$ *in the Algorithm 1, for the number of oracle calls equal to*

$$N = \left\lceil \frac{81 M_g^2 \Theta^2}{\varepsilon_g^2} \ln \frac{1}{\sigma} \right\rceil$$

the point $\bar{x}^N$ *satisfies*

$$\mathbb{P}\left\{|I| > 1, \quad f(\bar{x}^N) - f(x_*) \leqslant \varepsilon_f, \quad g(\bar{x}^N) \leqslant \varepsilon_g\right\} \geqslant 1 - \sigma.$$

Strongly Convex Case Algorithm

Algorithm 2 Restarting Mirror Descent

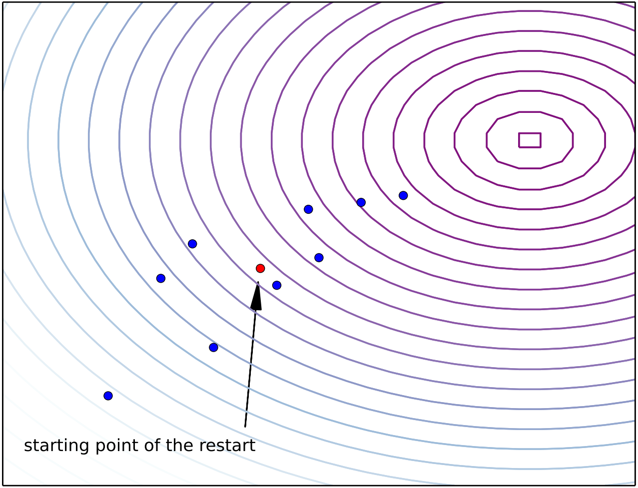
---

procedure $\textsc{RestartMirror}(x^0, N_1, \ldots, N_K, \Theta^2)$
    $\theta^2 := \Theta^2$
    for $k \in \{1, \ldots, K\}$ **do**
        $x^k \leftarrow \textsc{Mirror}(x^{k-1}, N_k, \theta^2)$
        $\theta^2 := \frac{1}{2}\theta^2$
        $d(x) := d(x - x^k + x^{k-1})$
    return x^K

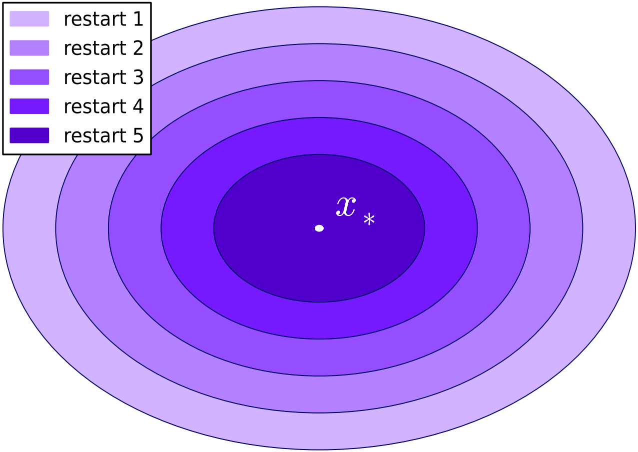
Intuition

- Each point returned by the $\textsc{Mirror}()$ procedure must be closer to the actual minimizer than the previous one
- The closer to minimizer we start, the faster we reach the required accuracy
- The point returned by the $\textsc{Mirror}()$ procedure is the average of step points, so by restarting we do not simply proceed iterating

Intuition



Intuition



Key Lemma

Lemma 1

Suppose f and g are μ-strongly convex functions with respect to the norm $\|\cdot\|$ over the convex set Q. Let

$$x_* = \arg\min_{x \in Q} \{f(x) : g(x) \leqslant 0\}.$$

Then if

$$f(x) - f(x_*) \leqslant \varepsilon_f, \quad g(x) \leqslant \varepsilon_g,$$

then

$$\frac{\mu}{2} \|x - x_*\|^2 \leqslant \max\{\varepsilon_f, \varepsilon_g\}.$$

Strongly Convex Case. Probability of Large Deviations

- It is sufficient to choose N_k in Algorithm 2 as

$$N_k = \left\lceil \frac{324 M^2 \omega_n \ln \bar{\sigma}^{-1}}{\mu^2 R_0^2} \, 2^k \right\rceil$$

- Here

$$\bar{\sigma} = \sigma \left(\log_2 \frac{\mu R_0^2}{2\varepsilon} \right)^{-1}$$

- The total number of restarts is

$$K = \left\lceil \log_2 \frac{\mu R_0^2}{2\varepsilon} \right\rceil$$

Strongly Convex Case. Probability of Large Deviations

Theorem 2

Suppose f and g are μ-strongly convex with respect to the norm $\|\cdot\|$. In the assumptions of the Theorem 1, with the total number of oracle calls equal to

$$N = \left\lceil \frac{324 M^2 \omega_n}{\mu\varepsilon} \Big(\ln \log_2 \frac{\mu R_0^2}{2\varepsilon} + \ln \frac{1}{\sigma} \Big) \right\rceil,$$

where

$$M = \max\{M_f, M_g\}, \quad \varepsilon = \max\{\varepsilon_f, \varepsilon_g\}, \quad R_0^2 = \max_{x,y \in Q} \{\|x - y\|^2\}$$

the point x^K, generated by the Algorithm 2, satisfies

$$\mathbb{P}\{ f(x^K) - f(x_*) \leqslant \varepsilon, \quad g(x^K) \leqslant \varepsilon \} \geqslant 1 - \sigma.$$

Summary

Results

- 'Restart' method is transferred to constrained case
- $O(\frac{1}{\mu\varepsilon})$ oracle calls
- Suitable for a non-euclidean setup

Outlook

- In non-euclidean setup the constant $\frac{M^2}{\mu}$ can be very large $\to$ composite problem statement

References

- Ben-Tal, A., Nemirovski, A.: Lectures on Modern Convex Optimization
- Bayandina, A., Gasnikov, A., Gasnikova, E., Matsievskiy, S.: Primal-dual Mirror Descent in Constrained Stochastic Optimization Problems. Submitted to Comp. Math. & Math. Phys. (2017)
- Juditsky, A., Nesterov, Yu.: Deterministic and Stochastic Primal-Dual Subgradient Algorithms for Uniformly Convex Minimization. Stochastic Systems. 4, 1, 44-80 (2014)