IE 598: Big Data Optimization **Fall 2016**

Lecture 15: From Subgradient Decent to Mirror Decent – October 13

Lecturer: Niao He *Scriber: Hongyi LI*

15.1 Recap

Recall that our goal is to solve the nonsmooth convex problem

$$\min_{x\in X} f(x)$$

where the objective function f is (possibly non-differentiable) convex, Lipschitz continuous, and the set X is convex and compact. We use the following notations to describe the Lipschitz continuity and diameter of the set:

$$\begin{cases} M = M_{\|\cdot\|_2}(f) := \sup_{x\in X} \frac{|f(x)-f(y)|}{\|x-y\|_2} \\ \Omega = \Omega(X) := \max_{x,y\in X} \frac{1}{2}\|x-y\|_2^2 \end{cases}$$

Subgradient Decent: at each iteration, runs a projection along the negative subgradient direction:

$$x_{t+1} = \Pi_X(x_t - \gamma_t g(x_t)) \tag{15.1}$$

Main results:

$$\min_{1\leq t\leq T} f(x_t) - f_* \leq \begin{cases} O(\frac{\sqrt{\Omega}M}{\sqrt{T}}) & \text{if } f \text{ is convex;} \\ O(\frac{M^2}{\mu T}) & \text{if } f \text{ is } \mu\text{-strongly convex} \end{cases}$$

Key Inequality (used to show the results):

$$\gamma_t g(x_t)^T(x_t - x_*) \leq \frac{1}{2}\|x_t - x_*\|_2^2 - \frac{1}{2}\|x_{t+1} - x_*\|_2^2 + \frac{\gamma_t^2}{2}\|g(x_t)\|_2^2 \tag{15.2}$$

15.2 Lower Bounds for Non-smooth Convex Optimization

While the convergence rates achieved by subgradient descent seems much worse than those achieved by gradient descent for smooth problems, we show below that in the worst case, one cannot improve the $O(1/\sqrt{t})$ and $O(1/t)$ rates for the convex and strongly convex situations, respectively, when using only block-box oriented methods that only have access to the subgradient of the objective function. The worst case function is given by the piecewise-linear function $f(x) = \max_{1\leq i\leq n} x_i$. We provide details below.

Theorem 15.1 *For any $1 \leq t \leq n$, $x_1 \in \mathbb{R}_n$,*
(1) there exists a M-Lipschitz continuous function f and a convex set X with diameter Ω, such that for any first-order method that generates:

$$x_t \in x_1 + span(g(x_1), ..., g(x_{t-1})), \ \ where\ g(x_i) \in \partial f(x_i), i = 1, \ldots, t-1$$

we always have

$$\min_{1\leq s\leq t} f(x_s) - f_* \geq \frac{\sqrt{\Omega}M}{2\sqrt{2}(1+\sqrt{t})}$$

(2) there exists a μ-strongly convex, M-Lipschitz continuous function f and a convex set X with diameter Ω, for any first-order method as described above, we always have

$$\min_{1\le s\le t} f(x_s) - f_* \ge \frac{M^2}{8\mu t}$$

Proof: (Nemirovski & Yudin 1979)
Let $X = \{x \in \mathbb{R}^n, \|x\|_2 \le R\}$ where $R = \sqrt{\frac{\Omega}{2}}$. Then $\frac{1}{2}\|x-y\|_2^2 \le \|x\|_2^2 + \|y\|_2^2 \le 2R^2$. Hence, $\Omega(X) \le \Omega$.
Let $f : \mathbb{R}^n \to \mathbb{R}$ s.t.

$$f(x) = C \cdot \max_{1\le i\le t} x_i + \frac{\mu}{2}\|x\|_2^2,$$

where C is some constant to be determined.

The subgradient of function f is given by,

$$\partial f(x) = \mu x + C \cdot conv\{e_i : i \text{ such that } x_i = \max_{1\le j\le t} x_j\}$$

Note that the optimal solution and optimal value of the porblem $\min_{x\in X} f(x)$ is given by

$$x_{*i} = \begin{cases} -\frac{C}{\mu t} & 1 \le i \le t \\ 0 & t < i \le n \end{cases} \quad \text{and } f_* = -\frac{C^2}{2\mu t}.$$

This is because, for $t < i \le n$, x_{*i}'s only increase $\frac{\mu}{2}\|x\|_2^2$, we can just set them all to 0; for $1 \le i \le t$, f is symmetric for all x_{*i}.

Without loss of generality, set $x_1 = 0$ and consider the worst subgradient oracle, that given an input x, it returns $g(x) = C \cdot e_i + \mu x$, with i being the first coordinate that $x_i = \max_{1\le j\le t} x_j$.

By induction, we can show that $x_t \in span(e_1, ..., e_{t-1})$. This implies for $1 \le s \le t$, $f(x_s) \ge 0$. Therefore

$$\min_{1\le s\le t} f(x_s) - f_* \ge \frac{C^2}{2\mu t}.$$

(1) Let $C = \frac{M\sqrt{t}}{1+\sqrt{t}}$, $\mu = \frac{M}{R(1+\sqrt{t})}$, then $\|\partial f(x)\|_2 \le C + \mu\|x\|_2 \le C + \mu R = M$. This implies that $f(x)$ is M-Lipschitz continuous. Moreover, we have

$$\min_{1\le s\le t} f(x_s) - f_* \ge \frac{C^2}{2\mu t} = \frac{M\sqrt{\Omega}}{2\sqrt{2}(1+\sqrt{t})}$$

(2) Let $C = \frac{M}{2}$, $\mu = \frac{M}{2R}$, then $\|\partial f(x)\|_2 \le C + \mu R = M$. This implies that $f(x)$ is M-Lipschitz continuous, μ-strongly convex.

$$\min_{1\le s\le t} f(x_s) - f_* \ge \frac{C^2}{2\mu t} = \frac{M^2}{8\mu t}$$

If the method returns $g(x)$ not be e_i where i is the smallest coordinate such that $x_i = \max_{1\le j\le t} x_j$, then we have $x_t \in span(e_{i_1}, ..., e_{i_{t-1}})$ for some $i_1, \ldots, i_{t-1}$, the situation is only getting worse; we still have $f(x_s) \ge 0, \forall s \le t$ and the analysis still hold. ■

Remark: To obtain an ϵ-solution, the number of subgradient call is $O(\frac{M^2\Omega}{\epsilon^2})$ for convex function, and $O(\frac{M^2}{\mu\epsilon})$ for strongly convex function. The above theorem indicates that these complexity bounds are indeed optimal.

Let us take a close look at the complexity bound for convex function,

$$O(\frac{M^2\Omega}{\epsilon^2}) = O\left(\frac{[M_{\|.\|_2}(f)\cdot \max_{x,y\in X}\|x-y\|_2]^2}{\epsilon^2}\right)$$

where the term $M_{\|.\|_2}(f)\cdot \max_{x,y\in X}\|x-y\|_2$ can be considered as the $\|.\|_2$-variation of f on $x\in X$.

15.3 Mirror Decent

Recall the subgradient decent updating rule can be equivalently written as

$$x_{t+1} = \operatorname*{argmin}_{x\in X}\{\frac{1}{2}\|x-x_t\|_2^2 + \langle \gamma_t g(x_t), x\rangle\} = \operatorname*{argmin}_{x\in X}\{\frac{1}{2}\|x\|_2^2 + \langle \gamma_t g(x_t) - x_t, x\rangle\}$$

Why should we use the Euclidean $\|.\|_2$ distance?
We will introduce a new algorithm, *Mirror Descent*, that generalizes subgradient descent with non-Euclidean distances.

15.3.1 Basic Setup

The Mirror Descent algorithm works as follows

$$x_{t+1} = \operatorname*{argmin}_{x\in X}\{V_\omega(x,x_t) + \langle \gamma_t g(x_t), x\rangle\} = \operatorname*{argmin}_{x\in X}\{\omega(x) + \langle \gamma_t g(x_t) - \nabla\omega(x_t), x\rangle\}$$

where

- **Bregman distance:** $V_\omega(x,y) = \omega(x) - \omega(y) - \nabla\omega(y)^T(x-y)$;
- **Distance generating function:** $\omega(x): X\to\mathbb{R}$ should be convex, continuously differentiable and 1-strongly convex with respect to some norm $\|.\|$, i.e. $\omega(x)\geq \omega(y) + \nabla\omega(y)^T(x-y) + \frac{1}{2}\|x-y\|^2$.

Note: Bregman distance is not a valid distance: $V_\omega(x,y)\neq V_\omega(y,x)$ and triangle inequality may not hold. This is often referred to as Bregman divergence. By definition, we always have $V_\omega(x,y)\geq \frac{1}{2}\|x-y\|^2$.

Given an input x and vector ξ, we will define the **prox mapping**:

$$\text{Prox}_x(\xi) = \operatorname*{argmin}_{u\in X}\{V_\omega(u,x) + \langle \xi, u\rangle\} \tag{15.3}$$

Mirror Descent update can be simplified as

$$x_{t+1} = \text{Prox}_{x_t}(\gamma_t g(x_t))$$

Example 1 (ℓ_2 setup): $X\subseteq\mathbb{R}^n$, $\omega(x) = \frac{1}{2}\|x\|_2^2$, $\|x\| = \|x\|_2$, then

(a) Bregman distance redues to $V_\omega(x,y) = \frac{1}{2}\|x-y\|_2^2$

(b) Prox-mapping reduces to $\text{Prox}_x(\xi) = \Pi_X(x-\xi)$

(c) Mirror decent reduces to subgradient decent.

Example 2 (ℓ_1 setup) : $X = \{x \in \mathbb{R}^n_+, \sum_{i=1}^n x_i = 1\}$, $\omega(x) = \sum_{i=1}^n x_i \ln(x_i)$, $\|x\| = \|x\|_1$. One can verify that $\omega(x)$ is 1-strongly convex with respect to the $\|\cdot\|_1$ norm on X. In this case, we have

(a) Bregman distance becomes to $V_\omega(x,y) = \sum_{i=1}^n x_i \ln(x_i/y_i)$, known as the Kullback-Leibler divergence.

(b) Prox-mapping becomes to

$$\text{Prox}_x(\xi) = \left(\sum_{i=1}^n x_i e^{-\xi_i}\right)^{-1} \begin{bmatrix} x_1 e^{-\xi_1} \\ \cdots \\ x_n e^{-\xi_n} \end{bmatrix}$$

(c) Mirror decent gives rise to multiplicative updates with normalization.

15.3.2 Convergence of Mirror ecent

We first present the useful three point identity lemma:

Lemma 15.2 (Three point identity) *For any $x, y, z \in dom(\omega)$:*

$$V_\omega(x,z) = V_\omega(x,y) + V_\omega(y,z) - \langle \nabla\omega(z) - \nabla\omega(y), x - y\rangle$$

Proof: This can be easily derived from definiton. We have

$$\begin{aligned} V_\omega(x,y) + V_\omega(y,z) &= \omega(x) - \omega(y) + \omega(y) - \omega(z) - \langle \nabla\omega(y), x-y\rangle - \langle \nabla\omega(z), y-z\rangle \\ &= V_\omega(x,z) + \langle \nabla\omega(z), x-z\rangle - \langle \nabla\omega(y), x-y\rangle - \langle \nabla\omega(z), y-z\rangle \\ &= V_\omega(x,z) + \langle \nabla\omega(z) - \nabla\omega(y), x-y\rangle \end{aligned}$$

■

Note: when $\omega(x) = \frac{1}{2}\|w\|_2^2$, this is the same as Law of cosines i.e.

$$\|z-x\|_2^2 = \|z-y\|_2^2 + \|y-x\|_2^2 + 2\langle z-y, y-x\rangle.$$

Theorem 15.3 *For mirror decent, let f be convex, then we have:*

$$\min_{1\le t\le T} f(x_t) - f_* \le \frac{V_\omega(x_*, x_1) + \frac{1}{2}\sum_{t=1}^T \gamma_t^2 \|g(x_t)\|_*^2}{\sum_{t=1}^T \gamma_t} \tag{15.4}$$

and

$$f\left(\frac{\sum_{t=1}^T \gamma_t x_t}{\sum_{t=1}^T \gamma_t}\right) - f_* \le \frac{V_\omega(x_*, x_1) + \frac{1}{2}\sum_{t=1}^T \gamma_t^2 \|g(x_t)\|_*^2}{\sum_{t=1}^T \gamma_t} \tag{15.5}$$

where $\|.\|_*$ denotes dual norm.

Proof: Since $x_{t+1} = \text{argmin}_{x\in X}\{\omega(x) + \langle \gamma_t g(x_t) - \nabla\omega(x_t), x\rangle\}$, by optimality condition, we have

$$\langle \nabla\omega(x_{t+1}) + \gamma_t g(x_t) - \nabla\omega(x_t), x - x_{t+1}\rangle \ge 0$$

From Three point identity, we have for $\forall x \in X$:

$$\langle \gamma_t g(x_t), x_{t+1} - x\rangle \le \langle \nabla\omega(x_{t+1}) - \nabla\omega(x_t), x - x_{t+1}\rangle = V_\omega(x, x_t) - V_\omega(x, x_{t+1}) - V_\omega(x_{t+1}, x_t)$$

$$\langle \gamma_t g(x_t), x_t - x\rangle \le V_\omega(x, x_t) - V_\omega(x, x_{t+1}) - V_\omega(x_{t+1}, x_t) + \langle \gamma_t g(x_t), x_t - x_{t+1}\rangle$$

By Young's inequality,

$$\langle \gamma_t g(x_t), x_t - x_{t+1}\rangle \le \frac{\gamma_t^2}{2}\|g(x_t)\|_*^2 + \frac{1}{2}\|x_t - x_{t+1}\|^2.$$

From the strongly convexity of $\omega(x)$, $V_\omega(x_{t+1}, x_t) \ge \frac{1}{2}\|x_t - x_{t+1}\|^2$. Adding these two inequalities, we get the key inequality:

$$\langle \gamma_t g(x_t), x_t - x_*\rangle \le V_\omega(x_*, x_t) - V_\omega(x_*, x_{t+1}) + \frac{\gamma_t^2}{2}\|g(x_t)\|_*^2 \tag{15.6}$$

By convexity, we have $\gamma_t(f(x_t) - f(x_*)) \le \langle \gamma_t g(x_t), x_t - x_*\rangle \le V_\omega(x_*, x_t) - V_\omega(x_*, x_{t+1}) + \frac{\gamma_t^2}{2}\|g(x_t)\|_*^2$. Taking summation leads to

$$\sum_{t=1}^{T} \gamma_t(f(x_t) - f(x_*)) \le V_\omega(x_*, x_1) + \frac{1}{2}\sum_{t=1}^{T}\gamma_t^2\|g(x_t)\|_*^2$$

Notice that

$$\sum_{t=1}^{T} \gamma_t(f(x_t) - f(x_*)) \ge \sum_{t=1}^{T}\gamma_t \min_{t=1,\ldots,T}(f(x_t) - f(x_*))$$

and that

$$\sum_{t=1}^{T} \gamma_t(f(x_t) - f(x_*)) \ge \sum_{t=1}^{T}\gamma_t \cdot [f(\hat{x}_T) - f(x_*)]$$

where $\hat{x}_T = \frac{\sum_{t=1}^{T}\gamma_t x_t}{\sum_{t=1}^{T}\gamma_t}$. we can see (15.4) and (15.5) hold ■

15.3.3 Mirror Decent vs. Subgradient Decent

Let f be convex, with proper choice of stepsize as before, we can see that the convergence of these two methods looks very similar.

- For Subgradient Decent, $\epsilon_t \sim O(\frac{\sqrt{\Omega_s}M_s}{\sqrt{t}})$,
  where $\Omega_s = \max_{x\in X}\frac{1}{2}\|x - x_1\|_2^2$ and $M_s = M_{\|\cdot\|_2}(f) = \max_{x\in X}\|g(x)\|_2$.
- For Mirror Decent, $\epsilon_t \sim O(\frac{\sqrt{\Omega_m}M_m}{\sqrt{t}})$,
  where $\Omega_m = \max_{x\in X} V_\omega(x, x_1)$ and $M_m = M_{\|\cdot\|}(f) = \max_{x\in X}\|g(x)\|_*$.

The rate remains the same, but the constants corresponding to Ω and M differ. In some cases, one can show that using Mirror Descent significantly improves upon the constant.

Case 1. Consider $X = \{x \in \mathbb{R}^n : x_i \ge 0, \sum_{i=1}^{n} x_i = 1\}$

- For subgradient decent, under ℓ_2 setup, $\omega(x) = \frac{1}{2}\|x\|_2^2$, $\|x\| = \|x\|_* = \|x\|_2$, we know $\Omega_s = 1$.
- For mirror decent, under ℓ_1 setup, $w(x) = \sum_{i=1}^{n} x_i \ln x_i$, $\|x\| = \|x\|_1, \|x\|_* = \|x\|_\infty$, if we choose $x_1 = \text{argmin}_{x\in X}\,\omega(x)$, then $\Omega_m \le \max_{x\in X}\omega(x) - \min_{x\in X}\omega(x) = 0 - (-\ln n) = \ln(n)$.

Therefore, the ratio between the efficiency estimates of GD and MD is

$$O\left(\frac{1}{\ln(n)} \cdot \frac{\max_{x\in X}\|g(x)\|_2}{\max_{x\in X}\|g(x)\|_\infty}\right)$$

Note that $\|g(x)\|_\infty \leq \|g(x)\|_2 \leq \sqrt{n}\|g(x)\|_\infty$. Hence, in the worst case, we can see mirror decent would be $O(\sqrt{n})$ faster than subgradient descent.

Case 2. Consider $X = \{x \in \mathbb{R}^n : x_i \geq 0, \|x\|_2 \leq 1\}$. In this case, for sub-gradient method, $\Omega_s = 1$; for mirror decent, $\Omega_m \leq O(\sqrt{n}\ln(n))$. Therefore, Ω_m is $\sqrt{n}$ larger than Ω_s, this offset the effect between M_m and M_s. In this case, mirror decent is not necessarily faster than subgradient decent.

References

[1] Anatoli Juditsky and Arkadi Nemirovski, "First Order Methods for Nonsmooth Convex Large-Scale Optimization, I: General Purpose Methods."

[2] A.S. Nemirovsky and D.B. Yudin, "Problem Complexity and Method Efficiency in Optimization," *SIAM Review Volume 27 Issue 2*, 1985, pp. 264.