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

Lecture 17: Smoothing Techniques II – October 20

Lecturer: Niao He *Scriber: Sai Kiran Burle*

Overview. In the last lecture, we discussed several smoothing techniques to approximate nonsmooth convex functions, particularly Nesterov's smoothing technique. In this lecture, we will drill down the algorithmic details of gradient descent when applied to solve the smooth counterpart and discuss its connection to proximal point algorithm.

17.0 Recap of Nesterov's Smoothing

Aside from Subgradient Descent or Mirror Descent, using smoothing technique is another way to address nonsmooth convex optimization problems:

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

Suppose function f can be represented as

$$f(x) = \sup_{y\in Y} \{\langle Ax+b, y\rangle - \phi(y)\}$$

where Y is a convex, compact set and ϕ is a convex function. Nesterov's smoothing of f is given by

$$f_\mu(x) = \sup_{y\in Y} \{\langle Ax+b, y\rangle - \phi(y) - \mu d(y)\} = (\phi + \mu d)^* (Ax+b)$$

where $\mu > 0$ is the smoothing parameter, function $d(.)$ is strongly convex, everywhere nonnegative and $\min_{y\in Y} d(y) = 0$.

From the last lecture, we show that when applying the accelerated gradient descent to solve the smoothing counterpart:

$$\min_{x\in X} f_\mu(x)$$

we arrive at

$$f(x_t) - f_* \le \underbrace{[f(x_t) - f_\mu(x_t)]}_{\text{approximation error}} + \underbrace{[f_\mu(x_t) - f_{\mu,x}]}_{\text{optimization error}} \le \mathcal{O}\left(\mu D_Y^2 + \frac{\|A\|_2^2 D_X^2}{\mu t^2}\right)$$

In order to achieve ϵ accuracy, one should choose $\mu = \mathcal{O}\left(\frac{\epsilon}{D_Y^2}\right)$, and this leads to the an overall iteration complexity $\mathcal{O}\left(\frac{\|A\|_2 D_X D_Y}{\epsilon}\right)$. Note that this is substantially better than the $O(1/\epsilon^2)$ complexity required by Subgradient Descent or Mirror Descent.

17.1 Some Examples

We mentioned that Nesterov's smoothing technique is more general and flexible comparing to other smoothing techniques, such as Moreau-Yosida regularization, Teboulle's smoothing based on recession function. We

provide below a simple example to illustrate the smoothed function under difference choices of proximity function $d(\cdot)$ and Fenchel representation.

Example: $f(x) = |x|$
Note that f admits the following two different representation:

$$f(x) = \sup_{|y|\le 1} yx$$
$$\text{OR} \qquad f(x) = \sup_{\substack{y_1,y_2\ge 0\\ y_1+y_2=1}} (y_1 - y_2)x$$

Here, $Y = \{y : |y| \le 1\}$ or $Y = \{y = (y_1, y_2) : y_1, y_2 \ge 0, y_1 + y_2 = 1\}$, function $\phi(y) \coloneqq 0$.

Now we consider different choices for the distance function $d(y)$.

1. $d(y) = \frac{1}{2}y^2$. Clearly, $d(.)$ is 1-strongly convex on $Y = \{y : |y| \le 1\}$, and $d(y) \ge 0$.
   Nesterov's smoothing gives rise to

$$f_\mu(x) = \sup_{|y|\le 1} \left\{yx - \frac{\mu}{2}y^2\right\} = \begin{cases} \frac{x^2}{2\mu}, & |x| \le \mu \\ |x| - \frac{\mu}{2}, & |x| > \mu \end{cases} \tag{17.1}$$

   which is the well-known Huber function.

   Remark. The same approximation can be obtained from Moreau-Yosida smoothing technique as follows:

$$f_\mu(x) = \inf_{y\in Y} \left\{|y| + \frac{1}{2\mu}\|y - x\|_2^2\right\}$$

2. $d(y) = 1 - \sqrt{1 - y^2}$. Clearly, $d(.)$ is 1-strongly convex on $Y = \{y : |y| \le 1\}$ and $d(y) \ge 0$.
   Nesterov's smoothing gives rise to

$$f_\mu(x) = \sup_{|y|\le 1} \left\{yx - \mu\left(1 - \sqrt{1 - y^2}\right)\right\} = \sqrt{x^2 + \mu^2} - \mu \tag{17.2}$$

   Remark. The same approximation can be obtained from Ben-Ta l&Teboulle's smoothing based on recession function:

$$|x| = \sup_y \{g(x + \mu) - g(y)\}, \quad g(y) = \sqrt{1 + y^2}$$
$$f_\mu(x) = \mu g\left(\frac{x}{\mu}\right) = \sqrt{x^2 + \mu^2}$$

3. $d(y) = y_1 \log y_1 + y_2 \log y_2 + \log 2$. Clearly, $d(.)$ is 1-strongly convex on $Y = \{(y_1, y_2) : y_1, y_2 \ge 0, y_1 + y_2 = 1\}$ and $d(y) \ge 0$.
   Nesterov's smoothing gives rise to

$$f_\mu(x) = \sup_{\substack{y_1,y_2\ge 0\\ y_1+y_2=1}} \{(y_1 - y_2)x - \mu(y_1 \log y_1 + y_2 \log y_2 + \log 2)\} = \mu \log\left(\frac{e^{-\frac{x}{\mu}} + e^{\frac{x}{\mu}}}{2}\right) \tag{17.3}$$

Remark. The same approximation can be obtained from Ben-Tal Teboulle smoothing based on recession function.

$$|x| = \max\{x, -x\} = \sup_y \{g(x+\mu) - g(y)\}, \quad g(y) = \log\left(e^{y_1} + e^{y_2}\right)$$
$$f_\mu(x) = \mu g\left(\frac{x}{\mu}\right) = \mu \log\left(e^{-\frac{x}{\mu}} + e^{\frac{x}{\mu}}\right)$$

17.2 Nesterov's smoothing and Moreau-Yosida regularization

In the previous example, we see that under the simple proximity function $d(y) = \frac{1}{2}\|y\|_2^2$, Nesterov's smoothing is equivalent to Yoreau-Yosida regularization. Indeed, this is true in general. Let f^* denote the conjugate of the function f. Suppose f is proper, convex and lower-semicontinuous, then

$$f(x) = \max_y \left\{y^T x - f^*(y)\right\}$$

Then we can show that

$$\begin{array}{rcll} f_\mu(x) & = & \max_y \left\{y^T x - f^*(y) - \frac{\mu}{2}\|y\|_2^2\right\} & (\textbf{Nesterov's smoothing}) \\ & = & \left(f^* + \frac{\mu}{2}\|\cdot\|_2^2\right)^*(x) & \\ & = & \inf_y \left\{f(y) + \frac{1}{2\mu}\|x-y\|_2^2\right\} & (\textbf{ Moreau-Yousida regularization}) \end{array} \tag{17.4}$$

where the last equation follows from the following Lemma 17.1(a).

Lemma 17.1 *Let f and g be two proper, convex and semi-continuous functions, then*

(a) $(f+g)^*(x) = \inf_y \{f^*(y) + g^*(x-y)\}$

(b) $(\alpha f)^*(x) = \alpha f^*\left(\frac{x}{\alpha}\right)$ *for* $\alpha > 0$

Proof:

(a) First, we prove the following equality

$$(f \square g)^*(x) = f^*(x) + g^*(x)$$

where $f \square g$ denotes the convlution operator,

$$f \square g = \inf_y \{f(y) + g(x-y)\}$$

$$\begin{aligned} (f\square g)^*(x) &= \sup_z \left\{z^T x - \inf_y \left(f(y) + g(z-y)\right)\right\} \\ &= \sup_z \left\{z^T x - \inf_{y_1+y_2=z} \left(f(y_1) + g(y_2)\right)\right\} \\ &= \sup_z \left\{\sup_{y_1+y_2=z} \left\{(y_1+y_2)^T x - f(y_1) - g(y_2)\right\}\right\} \\ &= \sup_{y_1,y_2} \left\{(y_1+y_2)^T x - f(y_1) - g(y_2)\right\} \\ &= \sup_{y_1} \left\{y_1^T x - f(y_1)\right\} + \sup_{y_2} \left\{{y_2}^T x - g(y_2)\right\} \\ &= f^*(x) + g^*(x) \end{aligned}$$

So, we get

$$(F\square G)^*(x) = F^*(x) + G^*(x)$$

Using $F = f^*$, and $G = g^*$,

$$(f^*\square g^*)^*(x) = f(x) + g(x)$$

Taking conjugate on both sides, we arrive at

$$(f^*\square g^*)(x) = (f+g)^*(x)$$

This leads to the desired result.

(b) This is because

$$\begin{aligned}(\alpha f)^*(x) &= \sup_y \left\{y^T x - \alpha f(y)\right\} \\ &= \alpha \sup_y \left\{y^T \left(\frac{x}{\alpha}\right) - f(y)\right\} \\ &= \alpha f^*\left(\frac{x}{\alpha}\right)\end{aligned}$$

■

17.3 Proximal Point Algorithms

Note that when computing the gradient of the smoothed function $\nabla f_\mu(x)$, for any smoothing forms used in (17.4), we will need to solve subproblems in the form

$$\min_y \left\{f(y) + \frac{1}{2}\|x-y\|_2^2\right\}.$$

The optimal solution to this subproblem is often referred to as the proximal operator, which shares many similarity as the projection operator we discussed earlier. We provide some basic results below.

17.3.1 Basics of Proximal Operators

Definition 17.2 *Given a convex function f, the* **proximal operator** *of f at a given point x is defined as*

$$prox_f(x) = \arg\min_y \left\{f(y) + \frac{1}{2}\|x-y\|_2^2\right\}$$

As an immediate observation, for any $\mu > 0$, we have

$$prox_{\mu f}(x) = \arg\min_y \left\{f(y) + \frac{1}{2\mu}\|x-y\|_2^2\right\}$$

Example: Let f be the indicator function of a convex set X, namely,

$$f(x) = \delta_X(x) = \begin{cases}0, & x \in X \\ +\infty, & x \notin X\end{cases}$$

Then the proximal operator reduces to the projection operator onto X, i.e.

$$prox_f(x) = \arg\min_{y \in X} \left\{ \frac{1}{2} \|x - y\|_2^2 \right\} = \Pi_X(x).$$

In general, the proximal operator possesses many similar properties as the projection operator as discussed earlier, e.g. treating optimal solution as fixed point, non-expansiveness, and decomposition.

Proposition 17.3 *Let f be a convex function, we have*

(a) **(Fixed Point)** *A point x_* minimizes $f(x)$ iff $x_* = prox_f(x_*)$.*

(b) **(Non-expansive)** $\|prox_f(x) - prox_f(y)\|_2 \leq \|x - y\|_2$.

(c) **(Moreau Decomposition)** *For any x, $x = prox_f(x) + prox_{f^*}(x)$.*

Proof:

(a) First, if x_* minimizes $f(x)$, we have $f(x) \geq f(x_*), \forall x \in X$. Hence,

$$f(x) + \frac{1}{2} \|x - x_*\|_2^2 \geq f(x_*) + \frac{1}{2} \|x_* - x_*\|_2^2$$

This implies that

$$x_* = \arg\min_x \left\{ f(x) + \frac{1}{2} \|x - x_*\|_2^2 \right\} = prox_f(x_*)$$

To prove the converse, consider if

$$x_* = prox_f(x_*) = \arg\min_x \left\{ f(x) + \frac{1}{2} \|x - x_*\|_2^2 \right\}$$

By the optimality condition, this implies that

$$0 \in \partial f(x_*) + (x_* - x_*) \implies 0 \in \partial f(x_*)$$

Therefore, x_* minimizes f.

(b) Let us denote $u_x = \text{prox}_f(x)$ and $u_y = \text{prox}_f(x_y)$. Equivalently,

$$x - u_x \in \partial f(u_x) \quad \text{and} \quad y - u_y \in \partial f(u_y).$$

Now we use the fact that ∂f is a monotone mapping, which tells us that

$$\langle x - u_x - (y - u_y), u_x - u_y \rangle \geq 0$$

Hence, we have

$$\langle x - y, u_x - u_y \rangle \geq \|u_x - u_y\|_2^2 .$$

By Cauchy Schwartz inequality, this leads to $\|u_x - u_y\|_2 \leq \|x - y\|_2$ as desired.

(c) Let $u = prox_f(x)$, or equivalently, $x - u \in \partial f(u)$. Note that we also have $u \in \partial f^*(x - u)$, this is equivalent to $x - u = prox_{f^*}(x)$. Hence, $x = u + (x - u) = prox_f(x) + prox_{f^*}(x)$.

■

17.3.2 Gradient Descent for Smoothed Function

Recall the definition of $f_\mu(x)$ in (17.4), we can see that the gradient is given by

$$\nabla f_\mu(x) = \frac{1}{\mu}(x - prox_{\mu f}(x)) \tag{17.5}$$

Since f_μ is $\frac{1}{\mu}$-smooth, gradient descent works as follows

$$x_{t+1} = x_t - \mu \nabla f_\mu(x_t)$$

From equation (17.5), this is equivalent as

$$x_{t+1} = prox_{\mu f}(x_t)$$

which is known as **proximal point algorithm**, initially proposed by Rockafellar in 1976. We discuss below the general algorithm and its convergence results.

17.3.3 Proximal Point Algorithm [RT76]

The goal is to minimize a non-smooth convex function $f(x)$, i.e. $\min_x f(x)$. The proximal point algorithm works as follows:

$$x_{t+1} = prox_{\gamma_t f}(x_t) \qquad t = 0, 1, 2, ...$$

where $\gamma_t > 0$ are the stepsizes.

Theorem 17.4 *Let f be a convex function, the proximal point algorithm satisfies*

$$f(x_t) - f_* \leq \frac{\|x_0 - x_*\|_2^2}{2\sum_{\tau=0}^{t-1} \gamma_\tau}$$

Proof: First, by optimality of x_{t+1}:

$$f(x_{t+1}) + \frac{1}{2\gamma_t}\|x_{t+1} - x_t\|_2^2 \leq f(x_t)$$

i.e.,

$$f(x_t) - f(x_{t+1}) \geq \frac{1}{2\gamma_t}\|x_{t+1} - x_t\|_2^2.$$

This further implies that $f(x_t)$ is non-increasing at each iteration. Note that this is essentially the key inequality when we analyze the convergence for Gradient Descent in Lecture 8. By far, we can simply adopt that analysis to get the desired result. To make it self-contained, we provide the full proof here.

Let $g \in \partial f(x_{t+1})$, by convexity of f, we have $f(x_{t+1}) - f_* \leq g^T(x_{t+1} - x_*)$. From the optimality condition of x_{t+1}, we have

$$0 \in \partial f(x_{t+1}) + \frac{1}{\gamma_t}(x_{t+1} - x_t) \implies \frac{x_t - x_{t+1}}{\gamma_t} \in \partial f(x_{t+1})$$

Hence,

$$\begin{aligned} f(x_{\tau+1}) - f_* &\leq \frac{1}{\gamma_\tau}(x_\tau - x_{\tau+1})^T(x_{\tau+1} - x_*) \\ &\leq \frac{1}{\gamma_t}(x_\tau - x_* + x_* - x_{\tau+1})^T(x_{\tau+1} - x_*) \\ &\leq \frac{1}{\gamma_t}\left[(x_\tau - x_*)^T(x_{\tau+1} - x_*) - \|x_{\tau+1} - x_*\|_2^2\right] \end{aligned}$$

Since $(x_\tau - x_*)^T(x_{\tau+1} - x_*) \leq \frac{1}{2}[\|x_\tau - x_*\|_2^2 + \|x_{\tau+1} - x_*\|_2^2]$, this implies that

$$\gamma_\tau(f(x_{\tau+1}) - f_*) \leq \frac{1}{2}\left[\|x_\tau - x_*\|_2^2 - \|x_{\tau+1} - x_*\|_2^2\right]$$

Summing this inequality for $\tau = 0, 1, 2, ...t-1$,

$$\sum_{\tau=0}^{t-1} \gamma_\tau(f(x_{\tau+1}) - f_*) \leq \frac{\|x_0 - x_*\|_2^2}{2} - \frac{\|x_t - x_*\|_2^2}{2} \leq \frac{\|x_0 - x_*\|_2^2}{2}$$

Since $f(x_\tau)$ is non-increasing, we have

$$\left(\sum_{\tau=0}^{t-1} \gamma_\tau\right)(f(x_t) - f_*) \leq \sum_{\tau=0}^{t-1} \gamma_\tau(f(x_{\tau+1}) - f_*)$$

Therefore,

$$f(x_t) - f_* \leq \frac{\|x_0 - x_*\|_2^2}{2\sum_{\tau=0}^{t-1} \gamma_\tau}$$

■

Remark. Note that

1. Unlike most algorithms we discussed so far in this course, the algorithm is not a gradient-based algorithm.
2. γ_t can be arbitrarily, the algorithm converges as long as $\sum_t \gamma_t \to \infty$, however, the cost of the proximal operator will depend on γ_t. For larger γ_t, the algorithm converges faster, but the proximal operator $prox_{\gamma_t f}(x_t)$ might be harder to solve.
3. If $\gamma_t = \mu$ (const.), then $f(x_t) - f_* = \mathcal{O}\left(\frac{1}{\mu t}\right)$. This matches with the $O(1/t)$ rate we obtain from the gradient descent perspective.

17.3.4 Accelerated Proximal Point Algorithm [GO92]

When combined with Nesterov's acceleration scheme, we get the **accelerated proximal point algorithm**, which works as follows

$$\begin{aligned} x_{t+1} &= prox_{\gamma_t f}(y_t) \\ y_{t+1} &= x_{t+1} + \beta_t(x_{t+1} - x_t) \end{aligned}$$

where β_t satisfies that

$$\beta_t = \frac{\alpha_t(1-\alpha_t)}{\alpha_t^2 + \alpha_{t+1}}$$
$$\alpha_{t+1}^2 = (\alpha_{t+1}+1)\alpha_t^2 - \frac{\mu}{L}\alpha_{t+1}$$

With the accelerated proximal point algorithm, the convergence can be improved to

$$f(x_t) - f_* \leq \mathcal{O}\left(\frac{1}{\left(\sum_{\tau=0}^{t-1}\sqrt{\gamma_t}\right)^2}\right)$$

When $\gamma_t = \mu$ (const.), then $f(x_t) - f_* = \mathcal{O}\left(\frac{1}{\mu t^2}\right)$. Again, this matches with the $O(1/t^2)$ rate we obtain from the gradient descent perspective.

References

[GO92] Güler, Osman. "New proximal point algorithms for convex minimization." *SIAM Journal on Optimization* 2.4 (1992): 649-664.

[RT76] Rockafellar, R. Tyrrell. "Monotone operators and the proximal point algorithm." *SIAM journal on control and optimization* 14.5 (1976): 877-898.