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

Lecture 16: Smoothing Techniques I – October 18

Lecturer: Niao He *Scribers: Harsh Gupta*

Overview. We discussed Subgradient Descent and Mirror Descent algorithms for non-smooth convex optimization in the past week. We observed that Subgradient Descent is a special case of the Mirror Descent algorithm. But, both these algorithms have general formulations and don't exploit the structure of the problem at hand. In practice, we always know some thing about the structure of the optimization problem we intend to solve. One can then utilize this structure to come up with more efficient algorithms as compared to Subgradient Descent and Mirror Descent algorithms.

16.1 Introduction

We intend to solve the following optimization problem:

$$\min_{x \in X} f(x) \tag{16.1}$$

where f is a convex but non-smooth, i.e., non-differentiable function, and X is a convex compact set. One intuitive way to approach the above problem is to approximate the non-smooth function $f(x)$ by a smooth and convex function $f_\mu(x)$, so that we can use the standard techniques learnt so far in the course to solve the problem. Hence, we want to reduce the problem in (16.1) to the following:

$$\min_{x \in X} f_\mu(x) \tag{16.2}$$

where $f_\mu(x)$ is a L_μ-Lipschitz continuous, smooth and convex approximation of the function $f(x)$. Now we can use the techniques learnt earlier in this course like gradient descent, accelerated gradient descent, Frank Wolfe algorithm, coordinate descent etc., to solve the above problem. Clearly, the objective now is to come up with a reasonably good approximation f_μ of f so that solving (16.2) is as close to solving (16.1) as possible.

A motivation example: Consider the simplest non-smooth and convex function, $f(x) = |x|$. The following function, known as the *Huber function*

$$f_\mu(x) = \begin{cases} \frac{x^2}{2\mu}, |x| \le \mu \\ |x| - \frac{\mu}{2}, |x| > \mu \end{cases} \tag{16.3}$$

is a smooth approximation of the absolute value function. We plot the two functions (for $\mu = 1$) in Figure 1. We make the following observations:

1. $f_\mu(x)$ is clearly continuous and differentiable everywhere. This can be seen straightforwardly from its formulation given in (16.3).
2. We observe that $f_\mu(x) \le f(x)$. Also, $f_\mu(x) \ge |x| - \frac{\mu}{2}$, therefore:

   $$f(x) - \frac{\mu}{2} \le f_\mu(x) \le f(x)$$

   Hence, if $\mu \to 0$, then $f_\mu(x) \to f(x)$. Therefore, μ characterizes the approximation accuracy.
3. We also observe that $|f''_\mu(x)| \le \frac{1}{\mu}$. This implies that $f_\mu(x)$ is $\frac{1}{\mu}$-Lipschitz continuous.

The Hubert function approximation has been widely used in machine learning to approximate non-smooth loss functions, e.g. absolute loss (robust regression), hinge loss (SVM), etc.

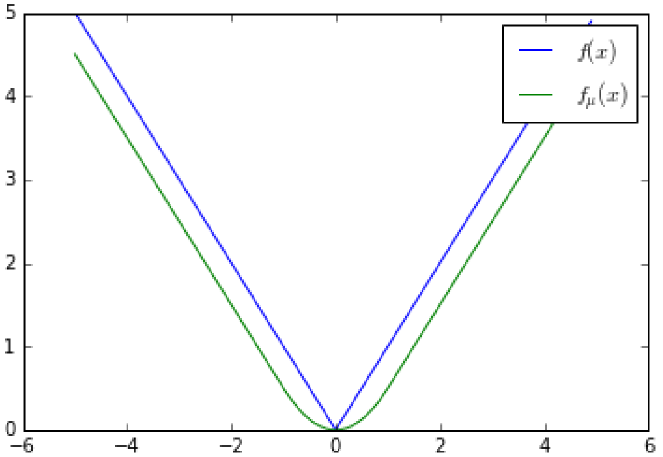

Figure 16.1: Plot for Example 1 ($\mu = 1$)

Robust Regression. Suppose we have m data samples $(a_1, b_1), ..., (a_m, b_m)$. We intend to solve the following regression problem with absolute loss:

$$\min_{x \in \mathbb{R}^d} \sum_{i=1}^{m} |{a_i}^T x - b_i|$$

We can approximate the absolute loss in the above optimization problem with the Huber loss and solve instead the following smooth convex optimization problem:

$$\min_{x \in \mathbb{R}^d} \sum_{i=1}^{m} f_\mu({a_i}^T x - b_i).$$

16.2 Smoothing Techniques

In this section, we will briefly introduce the major smoothing techniques used for non-smooth convex optimization.

1. Nesterov's Smoothing based on conjugate [Nesterov, 2005]

Nesterov's smoothing technique uses the following function to approximate $f(x)$:

$$f_\mu(x) = \max_{y \in \text{dom}(f^*)} \{x^T y - f^*(y) - \mu \cdot d(y)\} \tag{16.4}$$

where f^* is the convex conjugate of f defined as the following:

$$f^*(y) = \max_{x \in \text{dom}(f)} \{x^T y - f(x)\} \tag{16.5}$$

and $d(y)$ is some proximity function that is strongly convex and nonnegative everywhere.

2. Moreau-Yosida smoothing/regularization

Moreau-Yosida's smoothing technique uses the following function to approximate $f(x)$:

$$f_\mu(x) = \min_{y \in \text{dom}(f)} \{f(y) + \frac{1}{2\mu}||x - y||_M^2\} \tag{16.6}$$

where $\mu > 0$ is the approximation parameter, and the M-norm is defined as $||x||_M^2 = x^T M x$.

3. Ben-Tal-Teboulle smoothing based on recession function [Ben-Tal and Teboulle, 1989]

Ben-Tal and Teboulli's smoothing technique is applicable on a particular class of function which can be represented as following:

$$f(x) = F(f_1(x), f_2(x), ..., f_m(x)) \tag{16.7}$$

where $F(y) = \max_{x \in \mathrm{dom}(g)} \{g(x + y) - g(x)\}$ is the *recession function* of some function $g : \mathbb{R}^m \to \mathbb{R}$ here. For a function f satisfying the above condition, the Ben-Tal and Teboulli's smoothing technique uses the following function to approximate $f(x)$:

$$f_\mu(x) = \mu g \left(\frac{f_1(x)}{\mu}, ..., \frac{f_m(x)}{\mu} \right) \tag{16.8}$$

4. Randomized smoothing [Duchi, Bertlett, and Wainwright, 2012]

The randomized smoothing paradigm uses the following function to approximate $f(x)$:

$$f_\mu(x) = \mathbb{E}_Z f(x + \mu Z) \tag{16.9}$$

where Z is an isotopic Gaussian or uniform random variable.

In this lecture, we will mainly discuss Nesterov's smoothing and we will first discuss some conjugate theory in order to gain insight into this smoothing technique.

16.3 Conjugate Theory

Definition (Convex Conjugate). For any function $f : \mathbb{R}^n \to \mathbb{R}$, its convex conjugate is given as:

$$f^*(y) = \sup_{x \in \mathrm{dom}(f)} \{x^T y - f(x)\} \tag{16.10}$$

Note that f need not necessarily be convex for the above definition. Also, note that f^* will always be convex (regardless of f) since it is the supremum over linear functions of y.

By definition, we have:

$$f^*(y) \geq x^T y - f(x), \forall x, y \quad \Rightarrow \quad x^T y \leq f(x) + f^*(y), \forall x, y$$

The last inequality above is known as the Fenchel inequality. In the previous lectures, we studied the Young's inequality:

$$x^T y \leq \frac{||x||^2}{2} + \frac{||y||_*^2}{2}, \forall x, y$$

Note that the Young's inequality is essentially a special case of the Fenchel inequality. We now present the Moreau-Fenchel duality in the form of the following lemma.

Lemma 16.1 *If f is convex, lower semi-continuous and proper, then $(f^*)^* = f$.*

Here lower semi-continuity means that $\lim_{x \to x_0} \inf f(x) \geq f(x_0)$. In other words, the level set $(\{x : f(x) \leq \alpha\})$ of f is a closed set. A proper convex function means that $f(x) > -\infty$.

Thus for f satisfying the lemma, $f(x)$ admits the Fenchel representation

$$f(x) = \max_{y \in \mathrm{dom} f^*} \{y^T x - f^*(y)\}.$$

Proposition 16.2 *If f is μ-strongly convex then f^* is continuously differentiable and $\frac{1}{\mu}$-Lipschitz smooth.*

Proof:

By definition, we have $f^*(y) = \sup_{x \in \mathrm{dom}(f)} \{y^T x - f(x)\}$. This give us the subdifferential set

$$\partial f^*(y) = \arg \max_{x \in \mathrm{dom} f} \{y^T x - f(x)\}$$

Note that for all y, the optimal solution of the above problem is unique due to strong convexity. Hence, $\partial f^*(y)$ is a single set, i.e. $\partial f^*(y) = \nabla f^*(x)$. Hence, f^* is continuously differentiable. Now, we need to show the following:

$$|| \bigtriangledown f^*(y_1) - \bigtriangledown f^*(y_2)||_2 \le \frac{1}{\mu} ||y_1 - y_2||_2, \forall y_1, y_2 \tag{16.11}$$

Let $x_1 = \arg\max_{x \in \mathrm{dom} f} \{y_1^T x - f(x)\}$. Similarly, let $x_2 = \arg\max_{x \in \mathrm{dom} f} \{y_2^T x - f(x)\}$. From the optimality condition, we get:

$$\langle y_1, x_2 - x_1 \rangle \le \langle \partial f(x_1), x_2 - x_1 \rangle \tag{16.12}$$

$$\langle y_2, x_1 - x_2 \rangle \le \langle \partial f(x_2), x_1 - x_2 \rangle \tag{16.13}$$

From the μ-strong convexity of f, we have:

$$f(x_2) \ge f(x_1) + \partial f(x_1)^T (x_2 - x_1) + \frac{\mu}{2} ||x_2 - x_1||_2^2 \tag{16.14}$$

$$f(x_1) \ge f(x_2) + \partial f(x_2)^T (x_1 - x_2) + \frac{\mu}{2} ||x_1 - x_2||_2^2 \tag{16.15}$$

Combining equations (16.12), (16.13) with (16.14), (16.15) we get:

$$\langle y_1 - y_2, x_1 - x_2 \rangle \ge \mu ||x_1 - x_2||_2^2.$$

From the Cauchy-Schwarz inequality, this further implies that

$$\Rightarrow ||x_1 - x_2|| \le \frac{1}{\mu} ||y_1 - y_2||$$

Hence, (16.11) follows from the definitions of x_1, x_2. ■

Remark. Recall the Nesterov's smoothing technique

$$f_\mu(x) = \max_{y \in \mathrm{dom}(f^*)} \{x^T y - f^*(y) - \mu \cdot d(y)\} = (f^* + \mu d)^*(x)$$

by adding the strongly convex term $\mu d(y)$ term, the function $(f^* + \mu d)$ is strongly convex. Therefore, function $f_\mu(x)$ continuously differentiable and Lipschitz-smooth.

16.4 Nesterov's Smoothing

We consider a more generalized problem setting as compared to the previous sections. The goal is to solve the nonsmooth convex optimization problem

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

We assume that function f can be represented by

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

where $\phi(y)$ is a convex and continuous function and Y is a convex and compact set. Note that the aforementioned representation generalizes the Fenchel representation using conjugate function and needs not be unique. Indeed, for many cases, we are able to construct such representation easily as compared to using the convex conjugate.

Example. Let $f(x) = \max_{1\leq i\leq m} |a_i^T x - b_i|$. Computing the convex conjugate for f is a cumbersome task and f^* turns out to be very complex. But we can easily represent f as follows:

$$f(x) = \max_{y\in\mathbb{R}^m}\{(a_i^T x - b_i)y_i : \sum_i |y_i| \leq 1\}.$$

We now proceed to discuss some properties of the proximity function (or *prox function*) $d(y)$.

Proximity Function The function $d(y)$ should satisfy the following properties:

- $d(y)$ is continuous and 1-strongly convex on Y;
- $d(y_0) = 0$, for $y_0 \in \arg\min_{y\in Y} d(y)$;
- $d(y) \geq 0, \forall y \in Y$.

Let $y_0 \in Y$, here are some examples of valid prox functions:

- $d(y) = \frac{1}{2}||y - y_0||_2^2$
- $d(y) = \frac{1}{2}\sum w_i(y_i - (y_0)_i)^2$ with $w_i \geq 1$
- $d(y) = \omega(y) - \omega(y_0) - \nabla\omega(y_0)^T(y - y_0)$ with $\omega(x)$ being 1-strongly convex on Y

We can check that these proximity function also satisfies all the properties mentioned above.

Nesterov's smoothing considers the following smooth approximation of f

$$f_\mu(x) = \max_{y\in Y}\{\langle Ax + b, y\rangle - \phi(y) - \mu \cdot d(y)\}.$$

We first describe below the Lipschitz smoothness of this function [Nesterov, 2005].

Proposition 16.3 *For $f_\mu(x)$, we have*

- *$f_\mu(x)$ is continuously differentiable.*
- *$\bigtriangledown f_\mu(x) = A^T y(x)$, where $y(x) = \arg\max_{y\in Y}\{\langle Ax + b, y\rangle - \phi(y) - \mu \cdot d(y)\}$.*
- *$f_\mu(x)$ is $\frac{||A||_2^2}{\mu}$ Lipschitz smooth, where $\|A\|_2 := \max_x\{\|Ax\|_2 : \|x\|_2 \leq 1\}$.*

This can be derived similarly as Proposition 16.2, we omit the proofs here.

Now let us look at the approximation accuracy.

Theorem 16.4 *For any $\mu > 0$, let $D_Y^2 = \max_{y\in Y} d(y)$, we have*

$$f(x) - \mu D_Y^2 \leq f_\mu(x) \leq f(x).$$

This follows directly from the fact that $0 \leq d(y) \leq D_Y^2$.

Remark. Let $f_* = \min_{x\in X} f(x)$ and $f_{\mu,*} = \min_{x\in X} f_\mu(x)$, we have $f_{\mu,*} \leq f_*$. Moreover, for any x_t generated by an algorithm

$$f(x_t) - f_* \leq \underbrace{f(x_t) - f_\mu(x_t)}_{\text{approximation error}} + \underbrace{f_\mu(x_t) - f_{\mu,*}}_{\text{optimization error}}$$

Suppose we have access to compute the gradient of $f_\mu(x)$ when solving the resulting smooth convex optimization problem

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

(i) If we apply projected gradient descent to solve the smooth problem, then we have:

$$f(x_t) - f_* \leq O(\frac{||A||_2^2 D_X^2}{\mu t} + \mu D_Y^2)$$

Therefore, if we want the error to be less than a threshold ϵ, we need to set $\mu = O(\frac{\epsilon}{D_Y^2})$ and the total number of iterations is at most $T_\epsilon = O(\frac{||A||_2^2 D_X^2}{\epsilon\mu}) = O(\frac{||A||_2^2 D_X^2 D_Y^2}{\epsilon^2})$.

(ii) If we apply accelerated gradient descent to solve the smooth problem, then we have

$$f(x_t) - f_* \leq O(\frac{||A||_2^2 D_X^2}{\mu t^2} + \mu D_Y^2)$$

Therefore, if we want the error to be less than a threshold ϵ, we need to set $\mu = O(\frac{\epsilon}{D_Y^2})$ and the total number of iterations is at most $T_\epsilon = O(\frac{||A||_2 D_X}{\sqrt{\epsilon\mu}}) = O(\frac{||A||_2 D_X D_Y}{\epsilon})$.

In the latter case, the overall complexity $O(1/\epsilon)$ is substantially better than the $O(1/\epsilon^2)$ complexity when we directly apply subgradient descent to solve the original nonsmooth convex problem.

References

[Nes05] Y. Nesterov, “Smooth minimization of non-smooth functions,” *Math. Program.*,,Ser. A 103(1), 127–152 (2005)

[Ben89] A. Ben-Tal, and M. Teboulle, “A smoothing technique for nondifferentiable optimization problems.” In *Optimization*, pp. 1-11 (1989)

[Lem97] C. Lemaréchal and C. Sagastizábal, “ Practical aspects of the Moreau–Yosida regularization: Theoretical preliminaries”. SIAM Journal on Optimization, 7(2), 367-385 (1997)

[DBW12] J.C. Duchi, P.L. Bartlett, and M.J. Wainwright, “Randomized smoothing for stochastic optimization,” *SIAM J. OPTIM.*,,Vol. 22, No. 2, pp. 674–701 (2012)