EE 381V: Large Scale Optimization **Fall 2012**

Lecture 21 — November 13

Lecturer: Caramanis & Sanghavi *Scribe: Abhishek Kr Gupta*

21.1 Last Lecture

In the last lecture, we have studied stochastic subgradient methods and proved the convergence of noisy unbiased subgradient (NUS) method. We have also seen the randomized version of coordinate descent method and proved its convergence in expectation. Comparing it with the deterministic version, we saw that the random coordinate descent method becomes an attractive option once the ratio cost per iterate of the random method to the cost per iterate of the deterministic method becomes $O\left(\frac{1}{n}\right)$. Then this speed up in per iterate time will counter balance the slower linear convergence.

21.2 Introduction

Recall the unconstrained optimization problem

$$\min_x f(x).$$

We have seen the following results depending on the assumption for $f(x)$:

1. If $f(x)$ is not smooth (for example: L1 norm) which means $\nabla f(x)$ is not continuous, subgradient descent method gives the $\frac{1}{\sqrt{k}}$ convergence.
2. If $f(x)$ is smooth (*e.g.* $\nabla f(x)$ is L-Lipschitz continuous), gradient descent method gives the $\frac{1}{k}$ convergence.

There may be other methods (may be for particular functions) which can give better performance. Consider the following example:

$$f(x) \;=\; \frac{1}{2}\left\|Ax-b\right\|_2^2 + \left\|x\right\|_1.$$

This is a type 1 problem described above, as the second term is not smooth. However, the objective has a special form. It can be represented as sum of smooth and non -smooth function. Also note that second term $\left\|x\right\|_1$ is "simple" in a sense that we describe precisely below. In this case, it is conceivable that some tailored method might do better than the generic guarantees offered by sub gradient descent. Indeed, this is the case. In this lecture, we introduce what is known as the Proximal method, and show that it provides improved guarantees. In the next lecture, we discuss ways to build accelerated gradient methods for generic convex and smooth problems (*e.g.*, we show that for a generic convex smooth function it is possible to get $O(1/k^2)$, i.e., $O(1/\sqrt{\epsilon})$ convergence.

21.3 Proximal Map

Proximal method requires taking proximal map on each step which is defined as the following: The proximal mapping or prox-operator of a closed function $h(x)$ is defined as

$$\text{Prox}_h(x) = \arg\min_u h(u) + \frac{1}{2}\|u - x\|_2^2. \tag{21.1}$$

Note that $h(x)$ is a convex function so the above problem [21.1] is a convex problem, and in fact a strongly convex problem because of the added Euclidean norm term; therefore the solution is unique. Also for the proximal method to be efficient, we need that $h(x)$ should be simple in the sense that solving the problem [21.1] should be easier and fast or may be even analytical solutions are possible. This condition is required to ensure that each step is not much costly. We will now see some example functions and their proximal maps.

21.3.1 Examples:

Constant $h(x) = 0$

If $h(x) = 0$, the proximal map is given as

$$\begin{aligned}\text{Prox}_h(x) &= \arg\min_u h(u) + \frac{1}{2}\|u - x\|_2^2 \\ &= \arg\min_u \|u - x\|_2^2 = x,\end{aligned}$$

which is just identity map.

Convex Set Indicator function $h(x) = I_C(x)$

Recall the definition of an Indicator function is given as

$$I_C(x) = \begin{cases} 0 & x \in C \\ \infty & x \notin C \end{cases}.$$

The proximal map is given as

$$\begin{aligned}\text{Prox}_h(x) &= \arg\min_u I_C(x) + \frac{1}{2}\|u - x\|_2^2 \\ &= \arg\min_u \|u - x\|_2^2 \ , \ u \in C \\ &= \text{Proj}_C(x) = P_C(x)\end{aligned}$$

which is the same as projection over set C.

L1 Norm $h(x) = t\|x\|_1$

Now consider the case when $h(x)$ is L1 Norm. In that case, proximal map is given as

$$\text{Prox}_h(x) \quad = \quad \arg\min_u t\|x\|_1 + \frac{1}{2}\|u-x\|_2^2.$$

The above term is separable in indices. Each i^{th} coordinate can be optimized separately and the i^{th} coordinate is given as

$$\text{Prox}_h(x)_i \quad = \quad \text{sgn}(x_i)\left[|x_i| - \min\{|x_i|, t\}\right].$$

This operator is known as Soft thresholding operator and is shown in Figure 21.3.1.

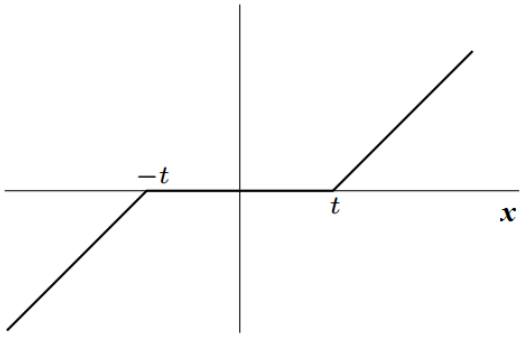

Figure 21.1. Soft Threshold Operator

21.3.2 Subgradient Characterization

We will now see some properties of proximal mapping from subgradient perspective. From optimality conditions of minimization in the definition, we can see that u^* is optimal point of $h(u) + \frac{1}{2}\|u-x\|_2^2$ only if $0 \in \partial h(x) + (u^* - x)$.

In other words,

$$\begin{aligned} u^* = \text{Prox}_h(x) \quad &\iff \quad 0 \in \partial h(x) + (u^* - x) \\ &\iff \quad x - u^* \in \partial h(x) \\ &\iff \quad h(z) = h(u) + (x-u)^T(z-u) \forall z. \end{aligned}$$

21.3.3 Properties

Proximal map has following properties:

Scaling

If $h = f(\lambda x + a)$, the proximal map of h can be given in terms of f as

$$\text{Prox}_h(x) = \frac{1}{\lambda}\left[\text{Prox}_{\lambda f}(\lambda x + a) - a\right].$$

Separable

If $h(x_1 + x_2) = h_1(x_1) + h_2(x_2)$,

$$\text{Prox}_h(x_1 + x_2) = \text{Prox}_{h_1}(x_1) + \text{Prox}_{h_2}(x_2).$$

Let us consider an example of L2 norm. let $h(x) = \|x\|_2^2$. Then proximal map is given as

$$\text{Prox}_{th}(x) = \frac{x}{\|x\|_2}(\|x\|_2 - t).$$

21.4 Proximal Gradient Method

We will now deduce a new optimization method based on computing proximal map at each update step. Consider the unconstrained optimization with objective split in two components

$$\min f(x) \text{ where } f(x) = g(x) + h(x),$$

where

- $g(x)$ convex, differentiable, smooth
- $h(x)$ closed convex with inexpensive prox-operator.

The update rule in the proximal gradient algorithm is given as

$$x^+ \;=\; \text{Prox}_{th}(x - t\nabla g(x)).$$

Here $t > 0$ is step size which can be either constant or determined by line search. Note that if $h(x) = 0$, then this method reduces to a gradient descent algorithm. Also if $h(x) = I_C(x)$, then this becomes projected gradient algorithm. [Refer to Sections above]

21.4.1 Interpretation

If we apply the definition of a proximal mapping to update rule, we get the following equation:

$$\begin{aligned} x_+ &= \arg\min_u th(u) + \frac{1}{2}\|u - x + t\nabla g(x)\|_2^2 \\ &= \arg\min_u h(u) + \frac{1}{2t}\|u - x + t\nabla g(x)\|_2^2 \\ &= \arg\min_u h(u) + g(x) + \nabla g(x)^T(u - x) + \frac{1}{2t}\|u - x\|_2^2. \end{aligned}$$

As we can note that second term is simple quadratic local model of $g(u)$ around x. So x_+ minimizes $h(u)$ plus this quadratic local model of $g(u)$ around x. Note that if $h = 0$, this will be just gradient descent. In that case update rule is given by

$$\begin{aligned} x_+ &= x - t\nabla g(x) \\ &= \arg\min_u g(x) + \nabla g(x)^T (u - x) + \frac{1}{2t} \|u - x\|_2^2. \end{aligned}$$

21.5 Convergence of Proximal gradient Method

In this section, we will see the convergence analysis of Proximal gradient method for fixed step t. We assume the following for the analysis:

- Step size is less than $\frac{1}{L}$, *i.e.* $t < \frac{1}{L}$ where L is the Lipschitz constant for gradient of g. In other words, we can say that

$$\|\nabla g(x) - \nabla g(y)| \leq L \|x - y\|.$$

- Optimum point exists and optimal function value $f^* > -\infty$.

We will first define gradient mapping and prove some inequalities.

21.5.1 Gradient Mapping

We can rewrite the update rule as following

$$\begin{aligned} x^+ &= \text{Prox}_{th}(x - t\nabla g(x)) \\ &= x - t\frac{1}{t}(x - \text{Prox}_{th}(x - t\nabla g(x))) \\ &= x - tG_t(x), \end{aligned}$$

where $G_t(x)$ is known as gradient mapping and defined as

$$G_t(x) = \frac{1}{t}(x - \text{Prox}_{th}(x - t\nabla g(x))).$$

Claim 1: $G_t(x) \in \nabla g(x) + \partial h(x - tG_t(x))$.

Proof: Note that from the definition,

$$\begin{aligned} G_t(x) &= \frac{1}{t}(x - \text{Prox}_{th}(x - t\nabla g(x))) \\ \Rightarrow x - tG_t(x) &= \text{Prox}_{th}(x - t\nabla g(x)). \end{aligned}$$

Now recall a basic property of the proximal mapping, that is essentially immediate from the definition:

$$u = \text{Prox}_h(x) \iff x - u \in \partial h(u).$$

Using this here, we get

$$\begin{aligned} x - t\nabla g(x) - x + tG_t(x) &\in \partial h(x - tG_t(x)) \\ \Rightarrow G_t(x) &\in \nabla g(x) + \partial h(x - tG_t(x)). \end{aligned}$$

□

Claim 2: If x^+ is the update as defined above, the following inequality holds for all z:

$$f(x^+) \leq f(z) - \frac{t}{2}\|Gt(x)\|_2^2 + (G_t(x))^T(x - z). \tag{21.2}$$

Proof: Recall from Lecture 4, for a convex function with L-Lipschitz gradient,

$$g(y) \leq g(x) - \nabla g(x)^T(y - x) + \frac{L}{2}\|y - x\|_2^2.$$

Let $y = x^+ = x - tGt(x)$, the above can written as

$$g(x - tGt(x)) \leq g(x) - t\nabla g(x)^T Gt(x) + t^2\frac{L}{2}\|Gt(x)\|_2^2.$$

For step size $t \leq 1/L$,

$$\begin{aligned} g(x - tGt(x)) &\leq g(x) - t\nabla g(x)^T Gt(x) + \frac{t}{2}\|Gt(x)\|_2^2 \\ \Rightarrow g(x^+) &\leq g(x) - t\nabla g(x)^T Gt(x) + \frac{t}{2}\|Gt(x)\|_2^2. \end{aligned} \tag{21.3}$$

Recall from the gradient map defination,

$$G_t(x) - \nabla g(x) \in \partial h(x - tG_t(x)) = \partial h(x^+).$$

From the definition of subgradient, we can say that

$$\begin{aligned} h(z) &\geq h(x^+) + (G_t(x) - \nabla g(x))^T(z - x^+) \\ \Rightarrow h(x^+) &\leq h(z) - (G_t(x) - \nabla g(x))^T(z - x^+). \end{aligned} \tag{21.4}$$

Adding the above two inequalities (21.3) and (21.4), we get,

$$\begin{aligned} g(x^+) &\leq g(x) - t\nabla g(x)^T Gt(x) + \frac{t}{2}\|Gt(x)\|_2^2 \\ h(x^+) &\leq h(z) - (G_t(x) - \nabla g(x))^T(z - x^+) \\ \Rightarrow g(x^+) + h(x^+) &\leq g(x) - t\nabla g(x)^T Gt(x) + \frac{t}{2}\|Gt(x)\|_2^2 \\ &\quad + h(z) + (G_t(x) - \nabla g(x))^T(x^+ - z) \\ \Rightarrow g(x^+) + h(x^+) &\leq g(z) + \nabla g(x)^T(x - z) - t\nabla g(x)^T Gt(x) \\ &\quad + \frac{t}{2}\|Gt(x)\|_2^2 + h(z) + (G_t(x) - \nabla g(x))^T(x^+ - z) \\ \Rightarrow g(x^+) + h(x^+) &\leq g(z) + h(z) + \frac{t}{2}\|Gt(x)\|_2^2 + (G_t(x))^T(x^+ - z). \end{aligned}$$

We get the following inequality as claimed,

$$f(x^+) \;\leq\; f(z) - \frac{t}{2}\left\|Gt(x)\right\|_2^2 + (G_t(x))^T(x-z).$$

□

21.5.2 Descent Method

We first show that proximal method is a descent method. The inequality (21.2) is true for all z. So let $z = x$,

$$f(x^+) \;\leq\; f(x) - \frac{t}{2}\left\|Gt(x)\right\|_2^2.$$

So each update give a new x which decreases the function value, So this is a descent method.

21.5.3 Convergence

If we put $z = x^*$ in (21.2), we get

$$\begin{aligned} f(x^+) - f(x^*) &\leq -\frac{t}{2}\left\|Gt(x)\right\|_2^2 + (G_t(x))^T(x-x^*) \\ &= \frac{1}{2t}\left[\|x-x^*\|^2 - \|x-x^*-tG_t(x)\|^2\right] \\ &= \frac{1}{2t}\left[\|x-x^*\|_2^2 - \left\|x^+-x^*\right\|_2^2\right] \\ f(x^{(i)}) - f(x^*) &= \frac{1}{2t}\left[\left\|x^{(i-1)}-x^*\right\|_2^2 - \left\|x^{(i)}-x^*\right\|_2^2\right]. \end{aligned}$$

If we sum over all $i \leq k$, the above can be written as

$$\begin{aligned} \sum f(x^{(i)}) - f(x^*) &= \frac{1}{2t}\left[\left\|x^{(0)}-x^*\right\|_2^2 - \left\|x^{(k)}-x^*\right\|_2^2\right] \\ &\leq \frac{1}{2t}\left[\left\|x^{(0)}-x^*\right\|_2^2\right]. \end{aligned}$$

Since $f(x^{(i)})$ is a decreasing sequence,

$$f(x^{(k)}) - f(x^*) \;\leq\; \frac{1}{k}\sum f(x^{(i)}) - f(x^*) \leq \frac{1}{2tk}\left[\left\|x^{(0)}-x^*\right\|_2^2\right] = O(\frac{1}{k}),$$

which proves $O(\frac{1}{k})$ convergence.

21.5.4 Line Search

Recall that we have assumed that $t < \frac{1}{L}$ only at one place to prove the following inequality

$$g(x^+ \quad \leq \quad g(x) - t\nabla g(x)^T Gt(x) + \frac{t}{2}\left\|Gt(x)\right\|_2^2 \tag{21.5}$$

But we generally don't know about L. For that case, we can do a line search to find a t which satisfies the above inequality (21.5). We will start at $t = 1$ and decrease it with a factor of β.

So at each step, the step size taken t_k is always between $\frac{1}{L}$ and $\frac{\beta}{L}$. So

$$\begin{aligned} t_{min} &\geq \frac{\beta}{L} \\ f(x^{(k)}) - f(x^*) &\leq \frac{1}{2t_{min}k}\left[\left\|x^{(0)} - x^*\right\|_2^2\right] = O(\frac{1}{k}). \end{aligned}$$

Therefore the convergence still remains of the order $O(\frac{1}{k})$ for the line search method as well.