EE 381V: Large Scale Optimization **Fall 2014**

Lecture 16 — October 21

Lecturer: Sanghavi *Scribe: Jimmy Lin and Taewan Kim*

16.1 Recap

Include table of time complexity

16.2 Motivation for Proximal Gradient Algorithm

blabla

16.2.1 Iterative Shrinkage-Thresholding Algorithm (ISTA)

ISTA!!!

16.2.2 Motivation 2

blabla

16.3 Proximal Gradient Algorithm

Proximal gradient is an algorithm for unconstrained problems with a cost function which can be expressed sum of two functions.

$$\text{minimize } f(x) = g(x) + h(x) \tag{16.1}$$

$$g(x) : \textit{'nice'},\ \text{(ex) convex, L-lipschitz gradient}$$

$$h(x) : \textit{'special'},\ \text{(ex) convex, not smooth, } \mathbf{prox} \text{ is easy to calculate}$$

In the above equation, we used the term 'nice' and 'special' to express the characteristic of each function g and h. Specific conditions on g and h will be more clear in the theorem which will be provided later. And the following is a definition of prox function (proximal operator) which clarifies the meaning of **prox** in the condition of h.

Definition 1. (Prox function/ Proximal operator)
The prox function (or proximal operator) of a function $h(\cdot)$ is defined as:

$$\mathbf{prox}_{th}(v) = \arg\min_x \left(h(x) + \frac{1}{2t}\|x - v\|_2^2 \right) \tag{16.2}$$

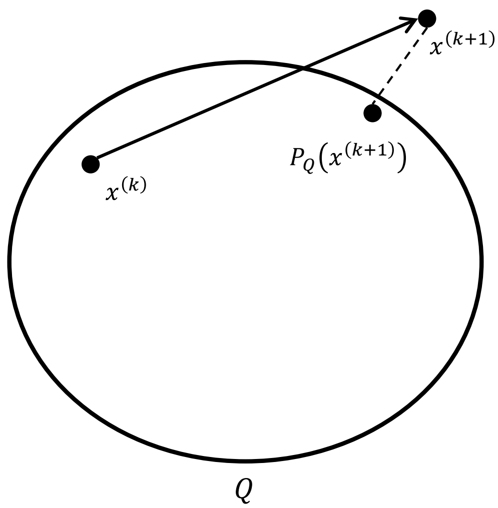

Figure 16.1. Projection on Q for gradient step

One natural example of proximal operator is projection on set Q. Suppose you perform a gradient descent and require a solution to be in a certain set Q. So, in each step, x_+ should satisfy the condition $x_+ \in Q$, and this can be done by using projection P_Q as in Figure 16.1.

$$x^{(k+1)} = P_Q\left(x^{(k)} - t\nabla g(x^{(k)})\right) \tag{16.3}$$

And this projection operator is a specific version of proximal operator by using $h(x) = I_Q(x)$ where $I_Q(\cdot)$ is a indicator function of Q.

$$\begin{aligned}
\mathbf{prox}_h(x) &= \arg\min_u \left(h(u) + \frac{1}{2}\|u-x\|_2^2\right) \\
&= \arg\min_u \left(I_Q(u) + \frac{1}{2}\|u-x\|_2^2\right) \\
&= \arg\min_{u\in Q} \left(\frac{1}{2}\|u-x\|_2^2\right) \\
&= P_Q(x)
\end{aligned}$$

where indicator function I_Q is defined as,

$$I_Q(x) = \begin{cases} 0 & \text{if } x \in Q \\ \infty & \text{if } x \notin Q \end{cases}$$

Third equality comes from the fact that $u \notin Q$ gives infeasible solution x for minimizing the function. And the last equality is the general definition of projection based on the Euclidean distance.

16.3.1 Proximal Gradient Algorithm

Now we can introduce a proximal gradient algorithm which uses two black boxes, one outputs $\nabla g(x)$ and the other outputs $\mathbf{prox}_{th}(v)$ (Figure 16.2), for the task of minimizing $f(x) = g(x) + h(x)$.

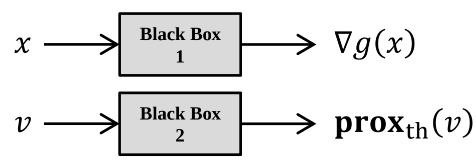

Figure 16.2. Using two black boxes for proximal gradient

- Proximal Gradient Algorithm

Proximal gradient algorithm is defined as the following update rule.

$$x_+ \leftarrow \mathbf{prox}_{th}\left(x - t\nabla g(x)\right) \tag{16.4}$$

By using the definition of proximal operator, update of proximal gradient can be expressed as follows:

$$\begin{aligned} x_+ &\leftarrow \mathbf{prox}_{th}\left(x - t\nabla g(x)\right) \\ &= \arg\min_u \left(h(u) + \frac{1}{2t}\|u - x + t\nabla g(x)\|_2^2\right) \\ &= \arg\min_u \left(h(u) + \langle \nabla g(x), u - x\rangle + \frac{1}{2t}\|u - x\|_2^2 + \frac{t}{2}\|\nabla g(x)\|_2^2\right) \\ &= \arg\min_u \left(h(u) + \langle \nabla g(x), u - x\rangle + \frac{1}{2t}\|u - x\|_2^2 + g(x)\right) \end{aligned} \tag{16.5}$$

Last equality comes from the fact that $\|\nabla g(x)\|_2^2$ and $g(x)$ do not depend on u. In equation (16.3.1), an important fact to point out is that rear part has a form of quadratic approximation of $g(u)$ around x.

$$g(x) + \langle \nabla g(x), u - x\rangle + \frac{1}{2t}\|u - x\|_2^2 \tag{16.6}$$

16.3.2 Convergence Analysis of Proximal Gradient

The convergence of proximal gradient algorithm requires $O(1/\epsilon)$ of iteration. Following theorem specifies the condition and time complexity.

Theorem 16.1. *If g is convex and g has L-lipschitz gradient, i.e. $\|\nabla g(x) - \nabla g(y)\|_2 \le L\|x - y\|_2$, using fixed size $t < 1/L$ on proximal gradient algorithm gives $O(1/\epsilon)$ of convergence to the optimal point x^*, where $x^* = \arg\min_x \left(g(x) + h(x)\right)$.*

Before providing a proof of the theorem, let's define a function $G_t(x)$ which satisfies the following condition. Role of the $G_t(x)$ function is to remove the proximal operator on the whole equation and make it similar to the general update as before.

$$x_+ \leftarrow x - tG_t(x) = \mathbf{prox}_{th}(x - t\nabla g(x))$$
$$\Leftrightarrow G_t(x) \triangleq \frac{1}{t}\left(x - \mathbf{prox}_{th}(x - t\nabla g(x))\right)$$

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

Lemma 16.2. *For any point z, $f(x_+) \leq f(z) + \langle G_t(x), x - z\rangle - \frac{t}{2}\|G_t(x)\|_2^2$*

Proof of Claim 1 and Lemma 16.2 is provided after the proof of Theorem 16.1.

Proof (Theorem 16.1): Putting $z = x$ in the Lemma 16.2 gives,

$$f(x_+) \leq f(x) - \frac{t}{2}\|G_t(x)\|_2^2$$

So, this gives the decreasing property $f(x_+) \leq f(x)$ since step size satisfies $t \geq 0$. And putting $z = x^*$ gives,

$$\begin{aligned} f(x_+) &\leq f^* - \frac{t}{2}\|G_t(x)\|_2^2 + \langle G_t(x), x - x^*\rangle \\ \Leftrightarrow f(x_+) - f^* &\leq \frac{1}{2t}\left[\|x - x^*\|_2^2 - \|x - x^* - tG_t(x)\|_2^2\right] \\ &= \frac{1}{2t}\left[\|x - x^*\|_2^2 - \|x_+ - x^*\|_2^2\right] \end{aligned}$$

Adding up over T iterations gives

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

From the fact that $f(x_+) \leq f(x)$ as shown above, $f(x^{(k)}) - f^* \geq f(x^{(T)}) - f^*$ for $k = 1, \cdots T$. And this gives,

$$f(x^{(T)}) - f^* \leq \frac{1}{2tT}\|x^{(0)} - x^*\|_2^2$$

which concludes the proof of $O(1/\epsilon)$ convergence, i.e. it requires $k = O(1/\epsilon)$ of iteration number to have $f(x^{(k)}) - f^* \leq \epsilon$.

□

Proof (Lemma 16.2): From the L-lipschitz continuous of gradient on g,

$$g(y) \leq g(x) + \langle \nabla g(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2$$

So, for $y = x_+ = x - tG_t(x)$

$$g(x_+) \leq g(x) - \langle \nabla g(x), tG_t(x) \rangle + \frac{L}{2}t^2\|G_t(x)\|_2^2 \tag{16.7}$$

And recall the condition of subgradient. If $a \in \partial h(x_+)$ then,

$$h(z) \geq h(x_+) + \langle a, z - x_+ \rangle$$

From the Claim 1, $G_t(x) - \nabla g(x) \in \partial h(x - tG_t(x))$. By setting $a = G_t(x) - \nabla g(x)$ gives,

$$\begin{aligned} & h(z) \geq h(x_+) + \langle G_t(x) - \nabla g(x), z - x_+ \rangle \\ \Leftrightarrow \;\; & h(x_+) \leq h(z) - \langle G_t(x) - \nabla g(x), z - x_+ \rangle \end{aligned} \tag{16.8}$$

Now using $f(x_+) = g(x_+) + h(x_+)$, and adding (16.7) and (16.8):

$$\begin{aligned} f(x_+) \leq \; & g(x) - \langle \nabla g(x), tG_t(x) \rangle + \frac{L}{2}t^2\|G_t(x)\|_2^2 \\ & + h(z) - \langle G_t(x) - \nabla g(x), z - x_+ \rangle \\ \leq \; & g(z) - \langle \nabla g(x), z - x \rangle - \langle \nabla g(x), tG_t(x) \rangle + \frac{L}{2}t^2\|G_t(x)\|_2^2 \\ & + h(z) - \langle G_t(x) - \nabla g(x), z - x_+ \rangle \\ & \qquad (\because g(x) \leq g(z) - \langle \nabla g(x), z - x \rangle \text{ from convexity of } g) \\ = \; & f(z) - \langle \nabla g(x), z - x + tG_t(x) - (z - x_+) \rangle \\ & + \frac{L}{2}t^2\|G_t(x)\|_2^2 - \langle G_t(x), z - x_+ \rangle \\ = \; & f(z) + \frac{L}{2}t^2\|G_t(x)\|_2^2 - \langle G_t(x), z - x + tG_t(x) \rangle \\ \leq \; & f(z) + \langle G_t(x), x - z \rangle + \frac{t}{2}\|G_t(x)\|_2^2 - t\|G_t(x)\|_2^2 \qquad (\because t < 1/L) \\ = \; & f(z) + \langle G_t(x), x - z \rangle - \frac{t}{2}\|G_t(x)\|_2^2 \end{aligned}$$

This concludes the proof of Lemma 16.2. □

Remaining part is the proof of Claim 1, i.e. $G_t(x) - \nabla g(x) \in \partial h(x - tG_t(x))$.

Proof (Claim 1): To show that $G_t(x) - \nabla g(x)$ is a subgradient of $h(x - tG_t(x)) = h(x_+)$, let's show that the following property of proximal operator is true.

$$u = \mathbf{prox}_{th}(x) \Leftrightarrow \frac{1}{t}(x - u) \in \partial h(u) \tag{16.9}$$

Proof of the property (16.9) is based on the definition of proximal operator, which means that $\mathbf{prox}_{th}(x)$ is a minimizer.

$$\begin{aligned}
&\frac{1}{t}(x-u) \in \partial h(u) \\
\Leftrightarrow\ & h(u) + \langle \frac{1}{t}(x-u), y-u \rangle \leq h(y), \quad \forall y \\
\Leftrightarrow\ & h(u) + \langle \frac{1}{t}(x-u), y-u \rangle - \frac{1}{2t}\|y-u\|_2^2 \leq h(y) - \frac{1}{2t}\|y-u\|_2^2 \leq h(y), \quad \forall y \\
\Leftrightarrow\ & h(u) + \langle \frac{1}{t}(x-u) + \frac{1}{2t}(u-y), y-u \rangle \leq h(y), \quad \forall y \\
\Leftrightarrow\ & h(u) + \frac{1}{2t}\langle 2(x-u) + (u-y), (x-u) - (x-y) \rangle \leq h(y), \quad \forall y \\
\Leftrightarrow\ & h(u) + \frac{1}{2t}\langle (x-u) + (x-y), (x-u) - (x-y) \rangle \leq h(y), \quad \forall y \\
\Leftrightarrow\ & h(u) + \frac{1}{2t}\left(\|u-x\|_2^2 - \|y-x\|_2^2\right) \leq h(y), \quad \forall y \\
\Leftrightarrow\ & h(u) + \frac{1}{2t}\|u-x\|_2^2 \leq h(y) + \frac{1}{2t}\|y-x\|_2^2, \quad \forall y \\
\Leftrightarrow\ & u = \arg\min_z \left(h(z) + \frac{1}{2t}\|z-x\|_2^2 \right) \\
\Leftrightarrow\ & u = \mathbf{prox}_{th}(x)
\end{aligned}$$

Now put $u = x_+ = \mathbf{prox}_{th}(x - \nabla g(x))$ in equation (16.9).

$$\begin{aligned}
\Rightarrow\ & \frac{1}{t}(x - \nabla g(x) - x_+) \in \partial h(x_+) \\
\Rightarrow\ & \frac{1}{t}(x - \nabla g(x) - x + tG_t(x)) \in \partial h(x_+) \\
\Rightarrow\ & G_t(x) - \nabla g(x) \in \partial h(x_+)
\end{aligned}$$

□