Large Scale Optimization: Lecture 16

Fall 2014

The University of Texas at Austin

Recap: Topics Covered in the previous lecture

Topics Covered by this Lecture

Proximal Gradient Algorithm

- Unconstrained problem with sum of two functions:

$$\text{minimize } f(x) = g(x) + h(x) \tag{1}$$

$g(x)$: *'nice'*, (ex) convex, L-lipschitz gradient

$h(x)$: *'special'*, (ex) convex, not smooth, **prox** is inexpensive

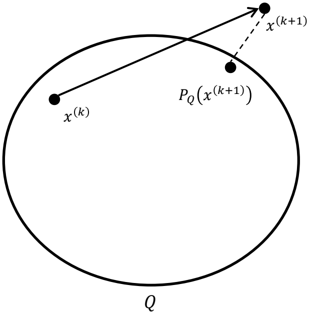
Definition

- **Prox function/ Proximal operator**

The prox function (or proximal operator) of a function $h(\cdot)$ is defined as:

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

Prox function/ Proximal operator



- Ex) Projection on set Q

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

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

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

Proximal Gradient Algorithm

- **Proximal Gradient Algorithm**

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

- Uses two black boxes for the task:

  minimize $f(x) = g(x) + h(x)$

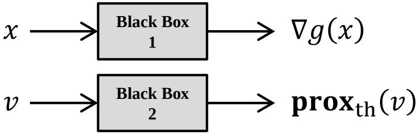


Figure: Using two black boxes for proximal gradient

Proximal Gradient Algorithm

- Form of quadratic approximation of $g(u)$ around x

$$
\begin{aligned}
x_+ \leftarrow & \mathbf{prox}_{th}(x - t\nabla g(x)) \\
&= \arg\min_u \left(h(u) + \frac{1}{2t}\|u - x + t\nabla g(x)\|_2^2 \right) \\
&= \arg\min_u \left(h(u) + \langle \nabla g(x), u - x\rangle + \frac{1}{2t}\|u - x\|_2^2 + \frac{t}{2}\|\nabla g(x)\|_2^2 \right) \\
&= \arg\min_u \left(h(u) + \langle \nabla g(x), u - x\rangle + \frac{1}{2t}\|u - x\|_2^2 + g(x) \right)
\end{aligned} \tag{3}
$$

Convergence Analysis of Proximal Gradient

Theorem

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

- Introducing $G_t(x)$

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

Convergence Analysis of Proximal Gradient - Preliminaries for proof

Claim

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

Lemma

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 and Lemma: Later

Proof of Theorem

Putting $z = x$ in the Lemma gives,

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

This gives the decreasing property $f(x_+) \leq f(x)$
And putting $z = x^*$ gives,

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

Proof of Theorem

Adding up over T iterations gives

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

From the fact that $f(x_+) \leq f(x)$,

$$f(x^{(k)})-f^* \geq f(x^{(T)})-f^* \text{ for } k=1,\cdots T$$

This gives,

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

$\Rightarrow$ $O(1/\varepsilon)$ convergence
(requires $k=O(1/\varepsilon)$ of iteration number for $f(x^{(k)})-f^* \leq \varepsilon$)

Proof of Lemma

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 $y = x_+ = x - tG_t(x)$

$$g(x_+) \leq g(x) - \langle \nabla g(x), tG_t(x) \rangle + \frac{L}{2}t^2\|G_t(x)\|_2^2 \tag{4}$$

Recall the definition of subgradient,

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

By Claim, $G_t(x) - \nabla g(x) \in \partial h(x - tG_t(x))$

$$\begin{aligned} & h(z) \geq h(x_+) + \langle G_t(x) - \nabla g(x), z - x_+ \rangle \\ \Leftrightarrow \;\; & h(x_+) \leq h(z) - \langle G_t(x) - \nabla g(x), z - x_+ \rangle \end{aligned} \tag{5}$$

Proof of Lemma

Add (4), (5) and use $f(x_+) = g(x_+) + h(x_+)$

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

Proof of Claim

Claim. $G_t(x) - \nabla g(x) \in \partial h(x_+)$

- Property of proximal operator

$$u = \mathbf{prox}_{th}(x) \Leftrightarrow \frac{1}{t}(x - u) \in \partial h(u) \tag{6}$$

(Proof: Use that $\mathbf{prox}_{th}(x)$ is a minimizer)

Now put $u = x_+ = \mathbf{prox}_{th}(x - \nabla g(x))$ in equation (6),

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

Next lecture

- Review of oracle complexity of algorithm
- Nesterov's accelerated gradient descent

Thank you!