Lecture 7 : Subdifferentiable Method

EE227C . Lecturer: Professor Martin Wainwright . Scribe: Alvin Wan

Let us now discuss subgradients. We desire the following subgradient method, for functions that are not differentiable but subdifferentiable.

$$x^{l+1} = x^l - \alpha g^l$$

where $g^l \in \partial f(x^l)$ and $\partial f(x) = \{g \in \mathbb{R}^d | f(y) \geq f(x) + \langle g, y - x \rangle\}, \forall y \in \mathbb{R}^d$. We're allowing this algorithm to pick any choice from the subdifferentiable. This is slightly different from taking the gradient. Here's an intuition for why we need to make a change right away.

Take this example, demonstrating that step size matters, $f(x) = |x|$.

$$\partial |x| = \begin{cases} +1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \\ [+1, -1] \text{ if } x = 0 \end{cases}$$

Suppose we were at $x^l = \frac{1}{2}$, then $g^l = 1$. However, imagine $\alpha = 1$, then we have $x^{l+1} = -\frac{1}{2}$, where $\alpha = 1, g^l = -1$, we will oscillate infinitely. Even with a smaller step size, we can play the same game. Unlike previously, no fixed α will work, so we will need time-varying step sizes. Let $\{\alpha^l\}$ be a sequence of step sizes that approaches 0. We showed the sequence of costs is non-increasing. Here, we will instead average our iterates and prove statements using sums. These sums will be built in a special way. In this case, we sum based on step sizes.

1 Proof for Subdifferentiable Method

Theorem: Say $C \subset \mathbb{R}^d$ which is closed, convex $f : \mathbb{R}^d \to \mathbb{R}$ and L-Lipschitz $\|f(x) - f(y)\| \leq L\|x - y\|_2, \forall x, y$. For instance, $f(x) = |x|$ above is 1-Lipschitz. This is true iff $\|g\|_2 \leq L, \forall g \in \partial f(x)$. Then,

$$f(\frac{\sum_{l=1}^{t} \alpha^l x^l}{\sum_{l=1}^{t} \alpha^l}) - f(x^*) \leq \frac{\|x^l - x^*\|_2^2}{2\sum_{l=1}^{t} \alpha^l} + \frac{L^2}{2} \frac{\sum_{l=1}^{t} (\alpha^l)^2}{\sum_{l=1}^{t} \alpha^l}$$

The first term in the bound considers the distance from x^l and x^*, divided by the running sum of step sizes. The second time considers the ratio between the sum of squares and the running sum. We want an infinite travel condition. This means as step sizes approach infinity, the steps should diverge. If not, our terms won't approach the optimal. The norm of the gradients is bounded, so consider: what is the largest amount we can ever move in space? It is the running sum of α^l. If the running sum doesn't diverge and approaches a finite sum, our adversary can start us some place farther than that finite sum from the optimal. We can achieve the convergence rate $\frac{1}{\sqrt{t}}$. So, let's talk about some step sizes.

Consider $\|x^l - x^*\|_2^2 \leq R^2$ and $\hat{X}^T = \frac{\sum_{l=1}^{t} \alpha^l x^l}{\sum_{l=1}^{t} \alpha^l}$.

1.1 Fixed iteration

Optimize for a fixed iteration number T. Fixed step size. $\alpha = \frac{R}{L\sqrt{T}}$. Then,

$$\hat{x}^T - f(x^*) \leq \frac{R^2}{2TR/L\sqrt{T}} + \frac{L^2}{2} \frac{TR^2/L^2T}{TR/L\sqrt{T}} = \frac{RL}{\sqrt{T}}$$

One is increasing and one is decreasing, so we picked a step size that balances the two terms. How would we set T? If we have some pre-prescribed tolerance ϵ, then simply solve $\frac{RL}{\sqrt{T}} = \epsilon$ to get $T = (\frac{RL}{\epsilon})^2$. If you know something about L, the quality of your starting point, and your tolerance for error, you can then compute the T. This algorithm is the slowest we've seen so far, but it's optimal.

1.2 General guarantee

For a general t guarantee, we pay a logarithmic price. Take $\alpha^l = \frac{R}{L\sqrt{l}}$, where $l = 1, 2, \ldots$. Then,

$$\sum_{l=1}^{t} \alpha^l \geq t\frac{R}{L\sqrt{t}} = \frac{R}{L}\sqrt{t}$$

What's interesting is that we don't have a finite sum $\sum_{l=1}^{T}(\alpha^l)^2$. This is diverging but only logarithmicly so.

$$\sum_{l=1}^{T}(\alpha^l)^2 \leq \frac{R^2}{L^2}(1+\log t)$$

Hence, $f(\hat{x}) - f(x^*) \leq \frac{RL(1+\log t)}{\sqrt{t}}$.

1.3 Generalized Update

Let's analyze $x^{l+1} = \Pi_C(x^l - \alpha^l g^l)$. Define $y^{l+1} = x^l - \alpha^l g^l, x^{l+1} = \Pi_C(y^{l+1})$. This is effectively the same proof we did for projected gradient methods, when we just had weak convexity. We will get a bound on the sub-optimality of x^l and then average using convexity. What's important is that $\frac{\alpha^l}{\sum_{l=1}^{t}\alpha^j}$ sums to 1, so we can apply Jensen's inequality. Now, by convexity,

$$f(x^l) - f(x^*) \leq \langle g^l, x^l - x^* \rangle = \frac{1}{\alpha^l}\langle x^l - y^{l+1}, x^l - x^* \rangle$$

Now, we use a polarization identity - a fancy name for an inequality of the form $\langle u, v\rangle = \frac{1}{2}(\|u\|_2^2 + \|v\|_2^2 - \|u-v\|_2^2)$.

$$f(x^l) - f(x^*) \leq \langle g^l, x^l - x^* \rangle \leq \frac{1}{2\alpha^l}(\|x^l - y^{l+1}\|_2^2 + \|x^l - x^*\|_2^2 - \|y^{l+1} - x^*\|_2^2)$$

Now, we have the following by non-convexity. In other words, negate both sides, and we get $-\|x^{l+1} - x^*\|_2 \geq -\|y^{l+1} - x^*\|_2$.

$$\|x^{l+1} - x^*\|_2 = \|\Pi_C(y^{l+1}) - \Pi_C(x^*)\|_2 \leq \|y^{l+1} - x^*\|_2$$

By our Lipschitz condition, we also have:

$$\|x^l - y^{l+1}\|_2^2 = (\alpha^l)^2 \|g^l\|_2^2 \leq (\alpha^l)^2 L^2$$

We've thus proved the following, by plugging in:

$$\alpha^l (f(x^l - x^*)) \leq \frac{1}{2}(\|x^l - x^*\|_2^2 - \|x^{l+1} - x^* \| 2^2) + \frac{(\alpha^l)^2 L^2}{2}$$

By Jensen's and convexity, we have our claim.

$$f(\sum_{l=1}^{T} \frac{\alpha^l}{\sum_{j=1}^{T} \alpha^j} x^l) - f(x^*) \leq \frac{1}{\sum_{j=1}^{T} \alpha^j} \sum_{l=1}^{T} \alpha^l (f(x^l) - f(x^*))$$

1.4 Danskin's Theorem

Take $f(x) = \max_{z \in Z} \phi(x, z)$ for $\phi : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ with $Z \subset \mathbb{R}^m$ compact and ϕ continuous. We have $x \to \phi(x, z)$ convex and $\nabla_x \phi(x, z)$ exists, continuous in z. Since this is convex, the subdifferentiable is non-empty. This theorem states that

$$\partial f(x) = \text{conv}(\nabla_x \phi(x, z^*) | z^* \in \text{argmax}_{z \in Z} \phi(x, z))$$

Example We can rewrite $f(x) = |x| = \max_{|z| \leq 1} xz$, where $\phi(x, z) = xz$. This gives us

$$\partial f(x) = \text{conv}(z^* | z^* \in \text{argmax}_{|z| \leq 1} \text{argmax} zx)$$

When is this argmax unique? When x is strictly non-zero. What if x is 0?

Example $f(x) = \|x\|_p, p \in (1, \infty)$

Example $f(x) = \|x\|_{op} = \max\{\sigma_1(x), \ldots, \sigma_d(x)\}$