Lecture 6 : Projected Gradient Descent

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

Consider the following update.

$$x^{l+1} = \Pi_C(x^l - \alpha \nabla f(x^l))$$

Theorem Say $f : \mathbb{R}^d \to \mathbb{R}$ is (m, M)-strongly convex smooth. Then with $\alpha = \frac{2}{m+M}$.

$$\|x^l - x^*\|_2 = (\frac{1 - m/M}{1 + m/M})^l \|x^0 - x^+\|_2$$

Take x^+ to be the unconstrained optimum, and the x^* to be the constrained optimum.

Proof: Define $S_\alpha(x) = x - \alpha \nabla f(x)$. Then with $\alpha = \frac{2}{m+M}$, we have

$$\|S_\alpha(x) - x^+\|_2 \leq \frac{1 - m/M}{1 + m/M} \|x - x^+\|_2$$

Define $S(x) = x - \alpha \nabla f(x)$ and $T(x) = T_C(S(x))$. Our method generates

$$x^{l+1} = \Pi_C(S(x^l)) = T(x^l)$$

In past lectures, we proved the following was a contraction, where $0 < \gamma < 1$. Note that

$$\|S(x) - S(y)\|_2 \leq \gamma \|x - y\|_2 \forall x, y \in \mathbb{R}^d$$

We proved this for $y = \text{argmin}_{u \in \mathbb{R}^d} f(u)$, but this holds in general. Now examine our algorithm. $\|T(x) - T(y)\|_2 = \|\Pi_C(S(x)) - \Pi_C(S(y))\|_2$ and $\|S(x) - S_(y)\| \leq \gamma \|x - y\|_2$ gives us that T is a γ contraction. This tells us rigorously that $T : \mathbb{R}^d \to C$ has a unique fixed point $\hat{x} \in C$, which satisfies $T(\hat{x}) = \hat{x}$. Unwrap our recursion to get the following.

$$\|x^l - \hat{x}\|_2 \leq \gamma \|x^0 - \hat{x}\|_2$$

We know that the following is approaching some value. However, we haven't used anything specific about the optimality conjecture - only non-expansiveness. So, we need to show that $\hat{x} = x^* = \text{argmin}_{x \in C} f(x)$. By optimality for projection, we have the following

$$\langle \Pi_C(\hat{x} - \alpha \nabla f(\hat{x})) - (\hat{x} - \alpha \nabla f(\hat{x})), y - \Pi_C(z) \rangle, \geq 0, \forall y \in C$$

We know $T(\hat{x}) = \Pi_C(\hat{x} - \alpha \nabla f(\hat{x})) = \hat{x}$, so the above becomes:

$$\alpha \langle \nabla f(\hat{x}), y - \hat{x} \rangle, \geq 0, \forall y \in C$$

How does this prove our claim? The gradient forms a positive quantity for all feasible directions. In other words, there is no other direction to descend along. Thus, $\hat{x} = x^*$.

Let's analyze f weakly convex, M-smooth. For unconstrained , we proved that with $\alpha = \frac{1}{M}$,

$$f(x^l) - f(x^*) \leq \frac{M \|x^0 - x^*\|_2^2}{2l}$$

Theorem Same bound holds for projected gradient descent. Assume M-smooth and weak convexity.

We claim the same bound with the same step size will hold in the projected case. We prove that the progress you make scales quadratically in the size of the gradient. In the unconstrained case, we showed

$$f(x^{l+1}) \leq f(x^l) - \frac{1}{2M} \|\nabla f(x^l)\|_2^2$$

We're no longer following the negative gradient. This algorithm is written as a descent method where the descent direction is different. Let us call it $g_C(x^l)$, and rewrite the algorithm that isolates g_c. We need to prove something analogous for constrianed problem. Consider $\alpha = \frac{1}{M}$

$$x^{l+1} = x^l - \alpha g_C(x^l)$$

Define $T(x) = \Pi_C(x - \frac{1}{M} \nabla f(x))$, then $g_C(x) = M(x - T(x))$. We've started with gradients, but there are plenty of other descent directions to use. We can replace the gradient with any direction that causes a potential function to decrease.

Lemma For $x, y \in C$, $f(T(x)) - f(y) \leq \langle f_C(x), x - y \rangle - \frac{1}{2M} \|g_C(x)\|_2^2$.

For an unconstrained problem, $g_c(X)$ is the gradient. Take lemma $u =^l, v = x^l$. We will get that

$$f(T(x^l)) - f(x^l) = f(x^{l+1}) - f(x^l) \leq -\frac{1}{2M}\|g_C(x^l)\|_2^2$$

The above is a descent guarantee. Compare this statement to what we did in the unconstrained case. Formally, it's the same. We can use the exact same algebra to get our telescoping property. Apply lemma with $u = x^l, v = x^*$.

$$\begin{aligned} f(x^{l+1}) - f(x^*) &\leq \langle f_C(x^l), x^l - x^* \rangle - \frac{1}{2M}\|g_C(x^l)\|_2^2 \\ &= \frac{M}{2}(\|x^l - x^*\|_2^2 - \|x^{l+1} - x^*\|_2^2) \end{aligned}$$

Now we can average.

$$\begin{aligned} f(x^T) - f(x^*) &\leq \frac{1}{T}\sum_{i=0}^{T-1}(f(x^{l+1}) - f(x^*)) \\ &\leq \frac{M}{2T}\sum_{l=0}^{T-1}(\|x^l - x^*\|_2^2 - \|x^{l+1} - x^+\|_2^2) \\ &\leq \frac{M}{2T}\|x^0 - x^*\|_2^2 \end{aligned}$$

If you look at this, it seems strange. We almost defined something circular, but it's actually the right intuition. Now, we prove the lemma above using closedness and convexity of C.

Proof of lemma: $S(x) = x - \frac{1}{M}\nabla f(x)$, and $T(x) = \Pi_C(S(x))$. By optimality conditions for projection, we have the following, where $T(x) = \Pi_C(S(x))$.

$$\forall y \in C, \langle T(x) - S(x), y - T(x) \rangle \geq 0 \leftrightarrow \langle \nabla f(x), T(x) - y \rangle \leq \langle g_C(x), T(x) - y \rangle$$

Then, we have $g_C(x) = M(x - T(x))$. For the first, we use M-smoothness. For the second term, we use convexity.

$$f(T(x)) - f(y) = (f(T(x)) - f(x)) + (f(x) - f(y))$$
$$\leq \langle \nabla f(x), T(x) - x \rangle + \frac{M}{2} \|\Pi_C(x) - x\|_2^2 + \langle \nabla f(x), x - y \rangle$$

Note that we have $\nabla f(x)x$ and $\nabla f(x)(-x)$ also note that we can relate the second term to g_C.

$$\langle \nabla f(x), T(x) - y \rangle + \frac{M}{2} \| \frac{1}{M} (g_C(x)) \|_2^2$$
$$\leq \langle \nabla f(x), T(x) - y \rangle + \frac{1}{2M} \|g_C(x)\|_2^2$$

Using the side result, we get the following.

$$\langle g_C(x), \Pi(x) - x + (x - y) \rangle + \frac{1}{2M} \|g_C(x)\|_2^2$$
$$= \langle g_C(x), x - y \rangle + (\frac{1}{2M} - \frac{1}{M}) \|g_C(x)\|_2^2$$

Intuitively, this algorithm works because of the direction $g_C(x)$ that works. M smoothness and convexity are very useful, powerful utilities. The overall conclusion is nontrivial, but when you use these simple tools correctly, you can build interesting things. So far, we've assumed functions are differentiable and convex. We will break down both of these. First, we relax the condition of differentiability.

Consider a sub-differentiable function. Why do I care? It's very clean and classical for convex functions. A standard problem that people like to solve is the following.

$$\min_{x \in \mathbb{R}^d} \|Ax - b\|_2^2 + \lambda \|x\|_1$$

Consider the matrix analog. Many researchers study matrix completion. Some are in particular interested in low-rank completion.

$$\min_{x \in S^{d \times d}} \sum_{(i,j) \in \Omega} (B_{ij} - X_{ij})^2 + \lambda \|X\|_{nuc}$$

This L1 norm is a weak surrogate for sparsity. In 2 dimensions, its epigraph is a convex set. Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$. Then, $z \in \mathbb{R}^d$ is a sub-gradient of f at x, written $z \in \partial f(x))$ if $f(y) \geq f(x) + \langle z, y - x \rangle \forall y \in \mathbb{R}^d$.

Generalizes $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$ for a convex, differentiable f. Geometrically, the gradient is any vector from the hyperplane and supporting the epigraph. When you're convex and differentiable, this is unique. subgradients are lines that we can draw through the point of discontinuity and not touch the epigraph. Consider the typical subdifferentiable function $|x|$.

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

Danskin's theorem gives us conditions on functions of the form:

$$f(x) = \max_{z \in \mathbb{Z}} \phi(x, z)$$

Under certain conditions, the subdifferentiable exists. The subdifferentiable is a convex hull of a set of vectors.

$$\partial f(x) = \text{conv}\{\nabla_x \phi(x, z) | \text{ z achieves} \max_z \phi(x, z)\}$$