10-725/36-725: Convex Optimization **Fall 2019**

Lecture 7: September 18

Lecturer: Ryan Tibshirani *Scribes: Yihang Shen, Allen Zhu, Mohsen Ferdosi*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

7.1 About the dual norm

The dual norm of $||\cdot||_2$ is $||\cdot||_2$, therefore:

$$f(x) = ||x||_2 = \max_{||z||_2 \leq 1} z^T x \tag{7.1}$$

When $x \neq 0$, we have:

$$\partial f(x) = \operatorname*{argmax}_{||z||_2 \leq 1} z^T x = \left\{ \frac{x}{||x||_2} \right\} \tag{7.2}$$

when $x = 0$, then $z^T x = 0$, we have:

$$\operatorname*{argmax}_{||z||_2 \leq 1} z^T x = z : ||z||_2 \leq 1 \tag{7.3}$$

7.2 Subgradient optimality conditions for lasso optimality

We want to characterize the set of lasso solutions

$$\min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1,\ \lambda > 0 \tag{7.4}$$

Subgradient optimality condition: 0 is in the subgradient of the objective

$$0 \in -X^T(y - X\beta) + \lambda\|\beta\|_1 \tag{7.5}$$

$$X^T(y - X\beta) = \lambda v,\ v \in \partial\|\beta\|_1 \tag{7.6}$$

7.3 Soft-thresholding

For lasso with $X = I$, we will have the solution $\beta = S_\lambda(y)$ such that:

$$[S_\lambda(y)]_i = \begin{cases} y_i - \lambda & \text{if } y_i > \lambda; \\ 0 & \text{if } -\lambda \leq y_i \leq \lambda; \\ y_i + \lambda & \text{if } y_i < \lambda; \end{cases} \tag{7.7}$$

Checking the subgradient optimality conditions:

When $y_i > \lambda$, $\beta_i = y_i - \lambda > 0$, so $y_i - \beta_i = \lambda = \lambda \text{Sign}(\beta_i)$.
Similarly for $y_i < \lambda$.
When $y_i \in [-\lambda, \lambda]$, $\beta_i = 0$, $|y_i - \beta_i| = |y_i| \leq \lambda$.

7.4 Distance to a convex set

Define the distance function to a closed, convex set C as:

$$\text{dist}(x, C) = \min_{y \in C} \|y - x\|_2 \tag{7.8}$$

This is well-defined with a unique solution since strongly convex ⇒ coercive ⇒ attains a unique minimum.

Then define the projection of x onto C as:

$$P_C(x) = \underset{y \in C}{\text{argmin}}\, ||y - x||_2^2 \tag{7.9}$$

Then we have:

$$\text{dist}(x, C) = \|x - P_C(x)\|_2 \tag{7.10}$$

If x is outside of C, then $\text{dist}(x, C) > 0$, we have:

$$\partial\, \text{dist}(x, C) = \left\{ \frac{x - P_C(x)}{\|x - P_C(x)\|_2} \right\} \tag{7.11}$$

Corollary: $\text{dist}(x, C)$ is differentiable for points outside C because the subgradient has a single element which is the gradient.

Proof of one direction is on the slide.

7.5 Convergence Analysis of Subgradient Method

This lecture we address optimization for non-differentiable f. Next lecture we speed up optimization using acceleration. That doesn't help for poorly conditioned problems; we will discuss second-order methods in a future lecture.

Subgradient methods behave differently from gradient descent. There is no notion of backtracking. Using fixed step sizes, they won't necessarily converge i.e. an optimality gap will remain even after infinite steps. We can prove convergence if we use diminishing step sizes which go to zero ($\sum t_k^2 < \infty$) but not too fast ($\sum t_k = \infty$), such as $t_k = 1/k$. In these cases, we can prove that $f(x_{\text{best}}^{(k)}) \to f^*$ at an $O(1/\epsilon)$ rate. This is a much weaker bound and it cannot be improved.

Theorem 7.1 *Assume that f convex, $dom(f) = \mathbb{R}^n$, and also that f is Lipschitz continuous with constant $G > 0$: $|f(x) - f(y)| \leq G\|x - y\|_2$ for all x, y. Then for a fixed step size t, subgradient method satisfies:*

$$\lim_{k \to \infty} f(x_{best}^{(k)}) \leq f^* + \frac{G^2 t}{2} \tag{7.12}$$

Where $f(x^{(k)}_{best}) = \min_{i=0,\dots,k} f(x^{(i)})$

Proof idea: bound the distance between each iterate and the optimal solution using the update rule $x^{(k)} = x^{(k-1)} - t^{(k)} g^{(k-1)}$.

Proof: We know that $(g^{k-1})^T(x^{k-1} - x^*) \geq f(x^{k-1}) - f^*$, then:

$$\begin{aligned} ||x^k - x^*||_2^2 &= ||x^{k-1} - tg^{k-1} - x^*||_2^2 && (7.13)\\ &= ||x^{k-1} - x^*||_2^2 - 2t(g^{k-1})^T(x^{k-1} - x^*) + t^2||g^{k-1}||_2^2 && (7.14)\\ &\leq ||x^{k-1} - x^*||_2^2 - 2t(f(x^{k-1}) - f(x^*)) + t^2||g^{k-1}||_2^2 && (7.15) \end{aligned}$$

Iterating, we have:

$$||x^k - x^*||_2^2 \leq ||x^0 - x^*||_2^2 - 2t\Sigma_{i=1}^k (f(x^{i-1}) - f(x^*)) + t^2 \Sigma_{i=1}^k ||g^{i-1}||_2^2 \qquad (7.16)$$

Note f being $G-$Lipschitz is equivalent to $||g||_2 \leq G$ so $||g^i||_2^2 \leq G^2$ for all i, and $||x^k - x^*||_2^2 \geq 0$, also define $R = \|x^0 - x^*\|_2$:

$$0 \leq R^2 - 2\sum_{i=1}^{k} t_i(f(x^{i-1}) - f(x^*)) + G^2 \sum_{i=1}^{k} t_i^2 \qquad (7.17)$$

Resulting in the *basic equation* from which we can read off all our convergence results:

$$f(x^k_{best}) - f(x^*) \leq \frac{R^2 + G^2 \sum_{i=1}^k t_i^2}{2\sum_{i=1}^k t_i} \qquad (7.18)$$

In particular, for fixed step size,

$$f(x^k_{best}) - f(x^*) \leq \frac{R^2}{2kt} + \frac{G^2 t}{2} \to \frac{G^2 t}{2} \text{ as } k \to \infty \qquad (7.19)$$

E.g. we can set $\dfrac{R^2}{2kt} = \dfrac{G^2 t}{2} = \dfrac{\epsilon}{2}$ by choosing $t = \epsilon/G^2$ and $k = R^2G^2/\epsilon^2$. Then the convergence rate is $O(1/\epsilon^2)$, which is much slower than gradient descent $O(1/\epsilon)$. We can't improve this convergence rate by varying step size or choosing unbalanced allocations. ■

7.6 Regularized logistic regression

Graph interpretations:

Ridge penalty $\|\beta\|_2^2$: gradient descends linearly. The criterion is strongly convex so it's easy to optimize.

Lasso penalty $\|\beta\|_1$: subgradient converges much, much more slowly.

7.7 Intersection of Sets with Polyak Step Size

Polyak step sizes: if the optimal criterion is known, we can choose t_k to derive the best convergence bound, which is still in $O(1/\epsilon^2)$. In most cases we don't know the optimal criterion, though we do for the intersection of sets. This is a cool example where the problem and solution don't seem to involve convex optimization.

Problem: We want to find a point in the intersection of closed, convex sets $C_1, \ldots, C_m$. We assume that we know how to project onto each set, but not the intersection. E.g. projecting onto halfspaces is easy but projecting onto their intersection (a polyhedron) is hard.

Let $f_i(x) = \text{dist}(x, C_i)$ for all $i = 1, \ldots, m$, and $f(x) = \max\limits_{i=1,\ldots,m} f_i(x)$. We want to solve $\min_x f(x)$ because $f^* = 0$ iff $x^* \in C_1 \cap \cdots \cap C_m$.

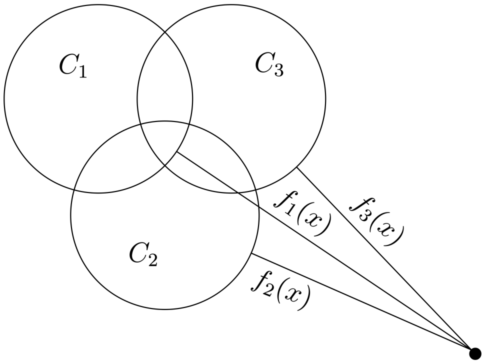

Figure 7.1: An illustration for $m = 3$

This is a convex problem because each f_i is convex, max is convex, and partial minimization over a convex set is convex.

The gradient of the distance function is:

$$\bigtriangledown \text{dist}(x, C) = \frac{x - P_C(x)}{||x - P_C(x)||_2} \tag{7.20}$$

And also we know that if $f(x) = \max_{i=1,\ldots,m} f_i(x)$, then:

$$\partial f(x) = \text{conv}(\cup_{i:f_i(x)=f(x)} \partial f_i(x)) \tag{7.21}$$

Set C_i furthest from x achieves the max, so $f_i(x) = f(x)$ and

$$g_i = \bigtriangledown f_i(x) = \frac{x - P_{C_i}(x)}{||x - P_{C_i}(x)||_2} \tag{7.22}$$

Then $g_i \in \partial f(x)$. We apply subgradient method with Polyak size $t_k = \frac{f(x^{k-1}) - f^*}{||g^{k-1}||_2^2} = f(x^{k-1})$:

$$\begin{aligned}
x^k &= x^{k-1} - f(x^{k-1}) \frac{x^{k-1} - P_{C_i}(x^{k-1})}{||x^{k-1} - P_{C_i}(x^{k-1})||_2} && (7.23)\\
&= x^{k-1} - f_i(x^{k-1}) \frac{x^{k-1} - P_{C_i}(x^{k-1})}{||x^{k-1} - P_{C_i}(x^{k-1})||_2} && (7.24)\\
&= x^{k-1} - ||x^{k-1} - P_{C_i}(x^{k-1})||_2 \frac{x^{k-1} - P_{C_i}(x^{k-1})}{||x^{k-1} - P_{C_i}(x^{k-1})||_2} && (7.25)\\
&= P_{C_i}(x^{k-1}) && (7.26)
\end{aligned}$$

So we've proved that we can find the optimum solution using the alternating projections algorithm. For the intersection of two sets, we just project back and forth between the two sets.