Lecture 7: September 21

Lecturer: Ryan Tibshirani *Scribes: Xiaoqi Chai, Ligong Han, Yang Zou*

7.1 Review of Subgradients

A subgradient of a convex function $f : \mathbb{R}^n \to \mathbb{R}$ is any value $g \in \mathbb{R}^n$ such that

$$f(y) \geq f(x) + g^T(y - x), \forall y$$

It always exists (on the relative interior of the domain).

7.2 Subgradient Optimality Condition

7.2.1 Subgradient Optimality Condition

Lemma 7.1 *For any function f (convex or not), x^* is a minimizer if and only if 0 is a subgradient of f at x^*:*

$$f(x^*) = \min_x f(x) \iff 0 \in \partial f(x^*)$$

Proof: $f(x^*) = \min_x f(x) \iff f(y) \geq f(x^*) \forall y \iff f(y) \geq f(x^*) + 0^T(y - x^*) \forall y \iff 0 \in \partial f(x^*)$ ∎

7.2.2 Derivation of First-Order Optimality Condition

If f is convex and differentiable, the subgradient optimality condition is equivalent to the first-order optimality condition.
Proof:

$$\begin{aligned} f(x^*) = \min_x f(x) &\iff f(x^*) = \min_x f(x) + I_C(x) \\ &\iff 0 \in \partial(f(x^*) + I_C(x^*)) \\ &\iff 0 \in \{\nabla f(x^*)\} + \mathcal{N}_C(x^*) \\ &\iff -\nabla f(x^*) \in \mathcal{N}_C(x^*) \\ &\iff -\nabla f(x^*)^T x^* \geq \nabla f(x^*)^T y, \text{ for all } y \in C \\ &\iff \nabla f(x^*)^T (y - x^*) \geq 0, \text{ for all } y \in C \end{aligned}$$

∎

7.2.3 Example 1: Lasso optimality conditions

Given a lasso problem

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

where $y \in \mathbb{R}^n, X \in \mathbb{R}^{n\times p}, \lambda \geq 0$, the subgradient optimality can be written as:

$$\begin{aligned} 0 \in \partial(\frac{1}{2}||y - X\beta||_2^2 + \lambda||\beta||_1) &\iff 0 \in \{-X^T(y - X\beta) + \lambda\partial||\beta||_1\} \\ &\iff X^T(y - X\beta) = \lambda v \\ &\iff \begin{cases} X_i^T(y - X\beta) = \lambda \text{sign}(\beta_i), & \text{if } \beta_i \neq 0 \\ |X_i^T(y - X\beta)| \leq \lambda, & \text{if } \beta_i = 0 \end{cases} \end{aligned}$$

where $v \in \partial||\beta||_1$

$$v_i \in \begin{cases} \{1\} & \text{if } \beta_i \geq 0 \\ \{-1\} & \text{if } \beta_i \leq 0, i = 1, \ldots, p \\ [-1, 1] & \text{if } \beta_i = 0 \end{cases}$$

- This provides a way to check lasso optimality

7.2.4 Example 2: Soft-thresholding

Consider the simpled lasso problem where $X = I$, from the example 1 the subgradient optimality conditions become:

$$\begin{cases} y_i - \beta_i = \lambda \text{sign}(\beta_i), & \text{if } \beta_i \neq 0 \\ |y_i - \beta_i \leq \lambda|, & \text{if } \beta_i = 0 \end{cases}$$

The solution can be solved from the optimality conditions. It is $\beta = S_\lambda(y)$, where $S_\lambda(y)$ is the soft-thresholding operator.

$$[S_\lambda(y)]_i = \begin{cases} y_i - \lambda & \text{if } y_i \geq \lambda \\ 0 & \text{if } -\lambda \leq y_i \leq \lambda, i = 1, \ldots, n \\ y_i + \lambda & \text{if } y_i \leq -\lambda \end{cases}$$

The plot of a soft-thresholding function is the following.

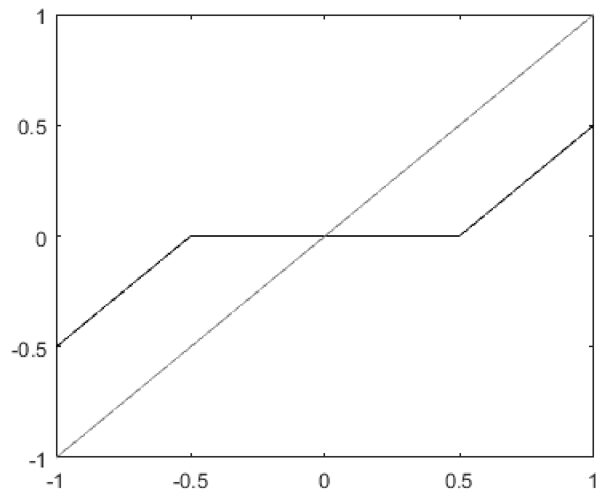

Figure 7.1: A soft-thresholding function

7.2.5 Example 3: Distance to a convex set

The distance function to a closed, convex set C is a convex function, which is:

$$\begin{aligned}\text{dist}(x, C) &= \min_{y \in C} ||y - x||_2 \\ &= ||x - P_C(x)||_2 \\ &\geq 0\end{aligned}$$

where $P_C(X)$, is the projection of x onto C.

- The subdifferential of the distance function $\partial\text{dist}(x, C)$ only has one element, so $\text{dist}(x, C)$ is differencitable and this is its gradient.

Proof: let $u = P_C(x)$.

$$\partial\text{dist}(x, C) = \{\frac{x - u}{||x - u||_2}\}$$

By the first-order optimality conditions,

$$(x - u)^T(y - u) \leq 0 \text{ for all } y \in C$$

$$C \subseteq H = \{y : (x - u)^T(y - u) \leq 0\}$$

(i) For $y \in H$,

$$(x - u)^T(y - u) \leq 0$$

$$\text{dist}(y, C) \geq 0$$

$$\text{dist}(y, C) \geq \frac{(x - u)^T(y - u)}{||x - u||_2}$$

(ii) For $y \notin H$, $(x-u)^T(y-u) = ||x-u||_2||y-u||_2 \cos\theta$, where θ is the angle between $x-u$ and $y-u$.

$$\frac{(x-u)^T(y-u)}{||x-u||_2} = ||y-u||_2 \cos\theta = \text{dist}(y,H) \leq \text{dist}(y,C)$$

Therefore, for any y,

$$\begin{aligned}\text{dist}(y,C) &\geq \frac{(x-u)^T(y-u)}{||x-u||_2} \\ &= \frac{(x-u)^T(y-x+x-u)}{||x-u||_2} \\ &= ||x-u||_2 + (\frac{x-u}{||x-u||_2})^T(y-x)\end{aligned}$$

Hence, $g = \frac{x-u}{||x-u||_2}$ is a subgradient of $\text{dist}(x,C)$ at x. ■

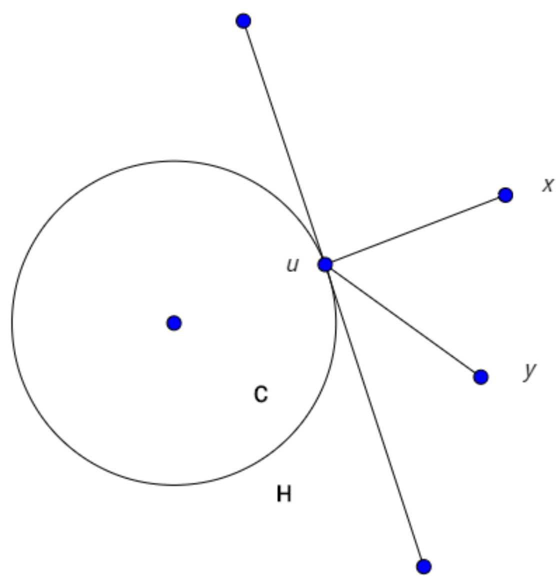

Figure 7.2: Diagram of the example 3

7.3 Subgradient Method

Like gradient descent, but replacing gradients with subgradients.

$$x^{(k)} = x^{(k-1)} - t_k \cdot g^{(k-1)}, k = 1,2,3,\ldots$$

where $g^{(k-1)} \in \partial f(x^{(k-1)})$, any subgradient of f at $x^{(k-1)}$

NOT necessarily descent!

7.3.1 Step size choices

- Fixed step sizes: $t_k = t$ all $k = 1, 2, 3, \dots$
- Diminishing step sizes:

$$\sum_{k=1}^{\infty} t_k^2 < \infty, \quad \sum_{k=1}^{\infty} t_k = \infty$$

  Aside: $\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}, \sum_{k=1}^{\infty} \frac{1}{k} = \infty$

7.3.2 Convergence analysis

Assume that f convex, $dom(f) = \mathcal{R}^n$, and also that f is Lipschitz with $G > 0$, i.e.

$$|f(x) - f(y)| \leq G\|x - y\|_2$$

for all x, y.

Theorem 7.2 *For a fixed step size t, subgradient method satisfies*

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

Theorem 7.3 *For diminishing step sizes, subgradient method satisfies*

$$\lim_{k \to \infty} f(x_{best}^{(k)}) = f^*$$

7.3.2.1 Converge rate

The basic inequality:

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

For fixed step sizes t,

$$f(x_{best}^{(k)}) - f^* \leq \frac{R^2}{2kt} + \frac{G^2 t}{2}$$

For this to be $\leq \epsilon$, choose $t = \epsilon / G^2$, and $k = R^2 G^2 / epsilon^2$ (converge rate $O(1/\epsilon^2)$, much slower than gradient descent $O(1/\epsilon)$)

7.3.2.2 Polyak step sizes

When the optimal value f^* is known, take

$$t_k = \frac{f(x^{(k-1)}) - f^*}{\|g^{(k-1)}\|_2^2}, k = 1, 2, 3, \dots$$

f^* can be estimated, gives same rate.

With Polyak step sizes, can show subgradient method converges to optimal value. Converge rate is still $O(1/\epsilon^2)$.

7.3.2.3 Can we do better?

Theorem 7.4 *(Nesterov): For any $k \leq n-1$ and starting point $x^{(0)}$, there is a function in the problem class such that any nonsmooth first-order method satisfies*

$$f(x^{(k)}) - f^* \geq \frac{RG}{2(1+\sqrt{k+1})}$$