10-725/36-725: Convex Optimization **Spring 2015**

Lecture 7: February 2

Lecturer: Ryan Tibshirani *Scribes: Mrinmaya Sachan*

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

This lecture's notes illustrate some uses of various LaTeX macros. Take a look at this and imitate.

7.1 Last Time

Last class, we learnt about gradient descent. Given an unconstrained convex optimization problem: $min\ f(x)$ where f is convex and differentiable, $dom(f) = \mathbb{R}^n$, the gradient descent algorithm started by choosing an initial point $x^{(0)} \in \mathbb{R}^n$ and then took successive steps to move along the direction of the negative gradient until convergence:

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

We also learnt about "backtracking" which is a popular approach for selecting the step size. If ∇f is Lipschitz, we also saw that the gradient descent algorithm had a convergence rate $\mathcal{O}(\frac{1}{\epsilon})$. While really simple and intuitive, the gradient descent algorithm requires the objective f to be differentiable. This limits the applicability of gradient descent. In this class we will study about subgradients, which is a generalization of gradients and can be used when the objective is non-differentiable.

7.2 Subgradients

Subgradients are counterpart of gradients for non-differentiable functions. They are closely related to the concept of convexity. Recall that a differentiable function f is said to be convex iff:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \qquad \forall x, y \in \mathbb{R}^n$$

In other words, a linear approximation always underestimates f. Subgradients are defined in a similar manner.

Definition 7.1 *A subgradient of a convex function f at a point x is any $g \in \mathbb{R}^n$ such that:*

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

While gradients may not exist/be undefined for non-differentiable function, subgradients always exist. If the function f is differentiable at the point x, then the subgradient g is unique and $g = \nabla f(x)$. This definition works for non-convex functions also. However, subgradients sometimes may not exist for non-convex functions.

7.2.1 Examples of Subgradients:

Example 1: $f(x) = |x|$. Wherever f is differentiable (i.e. $x \neq 0$), the subgradient is identical to the gradient, sign(x). At the point $x = 0$, anything in the range [1, 1] is a valid subgradient as any line passing through $x = 0$ with a slope in this range will lower bound the function.

Example 2: $f(x) = ||x||_2$. For $x \neq 0$, f is differentiable and the subgradient (also the gradient) is given by $g = \frac{x}{||x||_2}$. For x = 0, the subgradient is any vector whose 2 norm is at most 1. This holds because, by definition, in order for g to be a subgradient of f we must have $f(y) = ||y||_2 \geq f(x) + g^T(y - x) = g^T y \ \ \forall y$. In order for $||y||_2 \geq g^T y$ to hold, the only condition for g is $||g||_2 \leq 1$.

Example 3: $f(x) = ||x||_1, x \in \mathbb{R}^n$. Since $||x||_1 = \sum_{i=1}^n |x_i|$, we can consider each element g_i of the subgradient separately. The result is analogous to example 1. For $x_i = 0$, $g_i = sign(x_i)$. For $x_i = 0$, g_i is any point in [1, 1].

Example 4: $f(x) = max\{f_1(x), f_2(x)\}$, where f_1 and f_2 are arbitrary convex and differentiable functions. Here, we consider three cases. First, if $f_1(x) > f_2(x)$, then $f(x) = f_1(x)$ and therefore the subgradient is $g = \nabla f_1(x)$. Similarly, if $f_2(x) > f_1(x)$, then $f(x) = f_2(x)$ and $g = \nabla f_2(x)$. Finally, if $f_1(x) = f_2(x)$, then f may not be differentiable at x and anything on the line segment that joints $\nabla f_1(x)$ and $\nabla f_2(x)$ is a valid subgradient.

7.3 Subdifferential

Definition 7.2 *The subdifferential of a convex function $f : \mathbb{R}^n \to \mathbb{R}$ at some point x is the set of all subgradients of f at x:*

$$\partial f(x) = \{g : g \text{ is a subgradient of } f \text{ at } x\}$$

The subdifferential forms a closed and convex set. This holds even for non-convex functions. To verify this, suppose we have two subgradients $g_1, g_2 \in \partial f(x)$. We need to show that $g = \alpha g_1 + (1 - \alpha)g_2$ is also in $\partial f(x)$ for arbitrary $\alpha \in [0, 1]$. Using the definition of subgradients, we can write the following inequalities:

$$f(y) \geq f(x) + g_1^T(y - x) \ \ \forall y$$

$$f(y) \geq f(x) + g_2^T(y - x) \ \ \forall y$$

Now, multiplying the first with α and the second with $1 - \alpha$ and adding them up leads to the result:

$$f(y) \geq f(x) + \alpha g_1^T(y - x) + (1 - \alpha)g_2^T(y - x) = f(x) + g^T(y - x)$$

7.4 Connection to convex geometry

Consider the indicator function $I_C : \mathbb{R}^n \to \mathbb{R}$, defined on a convex set C:

$$\mathbb{I}_C(x) = \mathbb{I}(x \in C) = \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{if } x \notin C \end{cases}$$

For $x \in C$, $\partial \mathbb{I}_C(x) = \mathcal{N}_C(x)$, the normal cone of C at x. Recall that the normal cone of C at x is defined as:

$$\mathcal{N}_C(x) = \{g \in \mathbb{R}^n : g^T x \geq g^T y \text{ for any } y \in C\}$$

This directly follows from the definition of subgradient:

$$\mathbb{I}_C(y) \geq \mathbb{I}_C(x) + g^T(y-x) \quad \forall y$$

For $y \notin C$, $\mathbb{I}_C(y) = \infty$. For $y \in C$, $\mathbb{I}_C(y) = 0$. This means $0 \geq g^T(y-x)$ which is the definition of the normal cone.

The subgradients of indicator functions are important as any constrained coptimization problem $\min_{x \in C} f(x)$ can be rewritten as $\min_x f(x) + \mathbb{I}_C(x)$.

7.4.1 Subgradient Calculus

Subgradients for complex convex functions can be computed by knowing the subgradients for a basic set of convex functions and then applying rules of subgradient calculus. Here are the set of rules:

Scaling: $\partial(\alpha f) = \alpha \partial f$

Addition: $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$

Affine composition: If $g(x) = f(Ax) + b$ then $\partial g(x) = A^T \partial f(Ax + b)$.

Finite pointwise maximum: If $f(x) = \max_{i=1,\ldots,m} f_i(x)$ then $\partial f(x) = conv\left(\bigcup_{i:f_i(x)=f(x)} \partial f_i(x)\right)$

Generalized pointwise maximum: If $f(x) = \max_{x \in S} f_s(x)$ then $\partial f(x) \supseteq cl\left(conv\left(\bigcup_{s:f_s(x)=f(x)} \partial f_s(x)\right)\right)$

L_p **Norm:** $f(x) = ||x||_p$. Let q be such that $\frac{1}{p} + \frac{1}{q} = 1$, then, $||x||_p = \max_{||z||_q \leq 1} z^T x$. By duality theory, this leads to $\partial f(x) = \arg\max_{||z||_q \leq 1} z^T x = ||x||_q$

7.4.2 Why Subgradients?

Subgradients are important for two reasons. First, subgradients are useful for convex analysis - we can often give the optimality characterization via subgradients, monotonicity, relationship to duality. Second, if one can compute subgradients, then one can minimize (almost) any convex function.

7.4.2.1 Optimality condition

The optimality condition relates the minimizer of a function with the subgradient at the point. For any f (convex or not),

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

In other words, if x^* is a minimizer if and only if 0 is a subgradient of f at x. This is called the subgradient optimality condition. This is because $g = 0$ being a subgradient means that for all y

$$f(y) \geq f(x^*) + 0^T(y - x) = f(x^*)$$

7.4.2.2 Derivation of first order optimality

Recall that for f convex and differentiable, the problem $\min_{x \in C} f(x)$ is solved at x iff $\nabla f(x)^T(y-x) \geq 0 \ \ \forall y \in C$. In other words, the gradient increases as we move away from x. This can be seen by recasting the problem

as $\min_x f(x) + \mathbb{I}_C(x)$. Now, from the subgradient optimality condition $0 \in \partial(f(x) + \mathbb{I}_C(x))$. Also:

$$\begin{aligned} 0 \in \partial(f(x) + \mathbb{I}_C(x)) &\iff 0 \in \{\nabla f(x)\} + \mathcal{N}_C(x) \\ &\iff -\nabla f(x) \in \mathcal{N}_C(x) \\ &\iff -\nabla f(x)^T x \geq -\nabla f(x)^T y \ \ \forall y \in C \\ &\iff \nabla f(x)^T (y - x) \geq 0 \ \ \forall y \in C \end{aligned}$$

Example 1 - Lasso optimality conditions: Consider the following lasso problem ($y \in \mathbb{R}^n, X \in \mathbb{R}^{n \times p}$):

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}||y - X\beta||_2^2 + \lambda||\beta||_1$$

where, $\lambda \geq 0$. The subgradient optimality condition is:

$$\begin{aligned} 0 \in \partial\left(\frac{1}{2}||y - X\beta||_2^2 + \lambda||\beta||_1\right) &\iff 0 \in -X^T(y - X\beta) + \lambda\partial||\beta||_1 \\ &\iff X^T(y - X\beta) = \lambda v \end{aligned}$$

for some $v \in \partial||\beta||_1$. From Example 1 in Section 7.2.1,

$$v_i = \begin{cases} 1 & \text{if } \beta_i > 0 \\ -1 & \text{if } \beta_i < 0 \\ [-1, 1] & \text{if } \beta_i = 0 \end{cases}$$

Let $X_1, \ldots, X_p$ be the columns of X. Then, the subgradient optimality condition is:

$$\begin{cases} X_i^T(y - X\beta) = \lambda \ sign(\beta_i) & \text{if } \beta_i \neq 0 \\ |X_i^T(y - X\beta)| \leq \lambda & \text{if } \beta_i = 0 \end{cases}$$

Note that this is not an expression for a lasso solution. However, it does provide a way to check lasso optimality and to understand the lasso estimator: $|X_i^T(y - X\beta)| < \lambda$, then, $\beta_i = 0$.

Example 2 - Soft Thresholding: Consider the simplfied lasso problem with $X = I$:

$$\min_{\beta \in \mathbb{R}^n} \frac{1}{2}||y - \beta||_2^2 + \lambda||\beta||_1$$

We can solve this problem directly using subgradient optimality. The solution is $\beta = S_\lambda(y)$, where S_λ is the soft-thresholding operator:

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

Similar to the previous example, the subgradient optimality conditions are:

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

We can now plug in $\beta = S_\lambda(y)$ and check that these are satisfied:

- When $y_i > \lambda$, $\beta_i = y_i - \lambda > 0$, so $y_i - \beta_i = \lambda = \lambda 1$

- When $y_i < \lambda$, the argument is similar
- When $|y_i| \leq \lambda$, $\beta_i = 0$, and $|y_i - \beta_i| = |y_i| \leq \lambda$

Example 3 - Distance to a Convex set: Now, we consider the distance function to a convex set C:

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

It is easy to see that this is a convex function. Let us now derive its subgradients. We can rewrite $dist(x, C) = ||x - P_C(x)||_2$, where $P_C(x)$ is the projection of x onto C. Then when $dist(x, C) > 0$:

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

Since there is only one element in the subgradient, this implies that $dist(x, C)$ is indeed differentiable. We will only show one direction, i.e:

$$\frac{x - P_C(x)}{||x - P_C(x)||_2} \in \partial dist(x, C)$$

Let $u = P_C(x)$. Then, by first-order optimality conditions for a projection:

$$(u - x)^T (y - u) \geq 0 \ \ \forall y \in C$$

Hence,

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

Claim: for any y,

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

Note that for $y \in H$, the right-hand side is ≤ 0. For $y \notin H$, we have $(x-u)^T(y-u) = ||x-u||_2||y-u||_2 cos(\theta)$ where θ is the angle between $x - u$ and $y - u$. Thus:

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

as desired.

Using the claim, we have for any y,

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

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