10-725: Optimization **Fall 2012**

Lecture 6: Subgradient Method, September 13

Lecturer: Ryan Tibshirani *Scribes: Da-Cheng Juan, Jonathon M. Smereka, Kuan-Chieh Chen*

6.1 Intro to Subgradients

Some operations on convex functions destroy differentiability but preserve convexity - such as the max-operation. In these situations, subgradients offer a method of generalizing gradients for optimizing convex functions that are not necessarily differentiable (where gradient descent does not work).

6.1.1 Subgradients

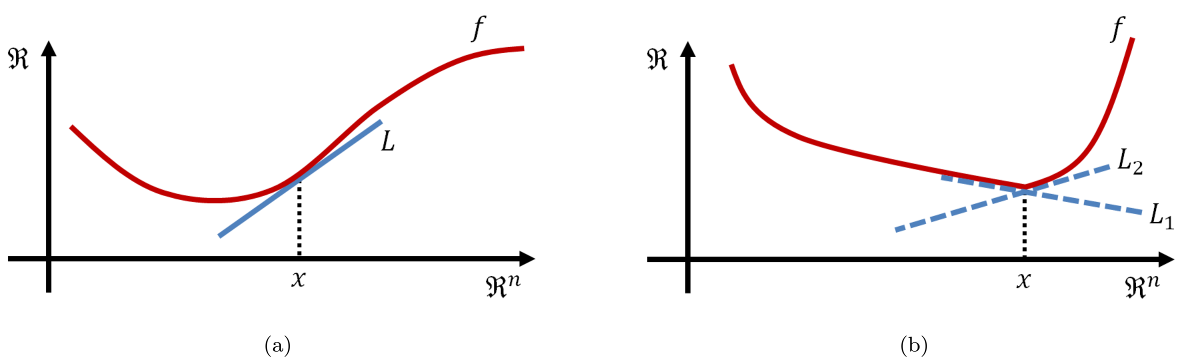

Figure 6.1

To say that a function $f : \Re^n \mapsto \Re$ is differentiable at x is to say that there is a single unique linear tangent such as shown in Fig 6.1a that under estimates the function:

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

While in Fig 6.1b we see the function f at x has many possible linear tangents that may fit appropriately. A **subgradient** is any $g \in \Re^n$ (same dimension as x) such that:

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

Thus, if a function is differentiable at a point x then it has a unique subgradient at that point ($\bigtriangledown f(x)$).

6.1.2 Subdifferentials

A **subdifferential** is the closed convex set of all subgradients of the convex function f:

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

Note that this set is guaranteed to be nonempty unless f is not convex.

6.1.3 Normal Cone

Often an indicator function, $I_C : \Re^n \mapsto \Re$, is employed to remove the contraints of an optimization problem (note that convex set $C \subseteq \Re^n$):

$$\min_{x \in C} f(x) \iff \min_x f(x) + I_C(x), \quad \text{where} \quad I_C(x) = I\{x \in C\} = \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{if } x \notin C \end{cases}$$

The subdifferential of the indicator function at x is known as the **normal cone**, $N_C(x)$, of C:

$$N_C(x) = \partial I_C(x) = \{g \in \Re^n : g^T x \geq g^T y \text{ for any } y \in C\}$$

6.2 Subgradient Calculus

Here, we provide some basic subgradient calculus for convex functions:

- Scaling: $\partial(af) = a \cdot \partial f$ provided $a > 0$. The condition $a > 0$ makes function f remain convex.
- 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 $f(x) = max_{i=1...m} f_i(x)$, then $\partial f(x) = conv\left(\bigcup_{i: f_i(x) = f(x)} \partial f_i(x)\right)$, which is the convex hull of union of subdifferentials of all active functions at x.
- General pointwise maximum: if $f(x) = max_{s \in S} f_s(x)$, then under some regularity conditions (on S, f_s), $\partial f(x) = cl\left\{conv\left(\bigcup_{s: f_s(x) = f(x)} \partial f_s(x)\right)\right\}$
- Norms: important special case, $f(x) = ||x||_p$. Let q be such that $1/p + 1/q = 1$, then $\partial f(x) = \left\{y : ||y||_q \leq 1 \text{ and } y^T x = \max_{||z||_q \leq 1} z^T x\right\}$
  Why is this a special case? Note $||x||_p = \max_{||z||_q \leq 1} z^T x$

6.3 Optimality condition

For a convex f,

$$f(x^*) = \min_{x \in \mathbb{R}^n} f(x) \Leftrightarrow 0 \in \partial f(x^*)$$

The reason is because $g = 0$ being a subgradient means that for all y

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

The analogy to the differentiable case is: $\partial f(x) = \{\nabla f(x)\}$.

6.4 Soft-thresholding

We use Lasso as an example to explain the concept of soft-thresholding. First, let us consider a simplified Lasso problem:

$$f(x) = \min_x \frac{1}{2}||y - x||^2 + \lambda||x||_1$$

And the solution of this problem is $x^* = S_\lambda(y)$, where $S_\lambda(y)$ is the soft-thresholding operator:

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

So the subgradients of $f(x)$ is

$$g = x - y + \lambda s,$$

where $s_i = sign(x_i)$ if $x_i \neq 0$ and $s_i \in [-1, 1]$ if $x_i = 0$. Now let $x^* = S_\lambda(y)$ and we can get $g = 0$. Why? If $y_i > \lambda$, we have $x_i^* - y_i = -\lambda + \lambda \cdot 1 = 0$. It is similar if $y_i < \lambda$. If $-\lambda \leq y_i \leq \lambda$, we have $x_i^* - y_i = -y_i + \lambda(\frac{y_i}{\lambda}) = 0$. Here, $s_i = \frac{y_i}{\lambda}$.

6.5 Subgradient method

Given a convex function $f : \mathbb{R}^n \to \mathbb{R}$, not necessarily differentiable. Subgradient method is just like gradient descent, but replacing gradients with subgradients. *I.e.*, initialize $x^{(0)}$, then repeat

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

where $g^{(k-1)}$ is **any** subgradient of f at $x^{(k-1)}$. We keep track of best iterate x_{best}^k among $x^{(1)}, \cdots, x^{(k)}$:

$$f(x_{best}^{(k)}) = \min_{i=1,\cdots,k} f(x^{(i)})$$

To update each $x^{(i)}$, there are basically two ways to select the step size:

- Fixed step size: $t_k = t$ for all $k = 1, 2, 3 \cdots$
- Diminishing step size: choose t_k to satisfy

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

6.6 Convergence analysis

Given the convex function $f : \mathbb{R}^n \to \mathbb{R}$ that satisfies:

- f is Lipschitz continuous with constant $G > 0$,

$$|f(x) - f(y)| \leq G||x - y|| \text{ for all } x, y$$

- $||x^{(1)} - x^*|| \leq R$ which means it is bounded

Theorem 6.1 *For a fixed step size t, subgradient method satisfies*

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

Proof:

$$\begin{aligned} ||x^{(k+1)} - x^*||^2 &= ||x^{(k)} - t_k g^{(k)} - x^*||^2 \\ &= ||x^{(k)} - x^*||^2 - 2t_k (g^{(k)})^T (x^{(k)} - x^*) + t_k^2 ||g^{(k)}||^2 \end{aligned}$$

By defintion of the subgradient method, we have

$$\begin{aligned} f(x^*) \geq f(x^{(k)}) + g^{(k)}(x^* - x^{(k)}) \\ -{g^{(k)}}^T \leq -(f(x^{(k)}) - f(x^*)) \end{aligned}$$

Using this inequality, we have

$$\begin{aligned} ||x^{(k+1)} - x^*||^2 &\leq ||x^{(k)} - x^*||^2 - 2t_k(f(x^{(k)}) - f(x^*)) + t_k||g^{(k)}||^2 \\ &\leq ||x^{(1)} - x^*||^2 - 2\sum_{i=1}^{k} t_i(f(x^{(i)}) - f(x^*)) + \sum_{i=1}^{k} t_i^2 ||g^{(i)}||^2 \end{aligned}$$

And this is lower bounded by 0, then we have

$$\begin{aligned} 0 \leq ||x^{(k+1)} - x^*||^2 \leq R^2 - 2\sum_{i=1}^{k} t_i(f(x_{(i)}) - f(x^*)) + \sum_{i=1}^{k} t_i^2 G^2 \\ 2\sum_{i=1}^{k} t_i(f(x^{(i)}) - f(x^*)) \leq R^2 + \sum_{i=1}^{k} t_i^2 G^2 \\ 2(\sum_{i=1}^{k} t_i)(f(x^{(k)}_{best}) - f(x^*)) \leq R^2 + \sum_{i=1}^{k} t_i^2 G^2 \end{aligned}$$

For a constant step size $t_i = t$:

$$\frac{R^2 + G^2 t^2 k}{2tk} \to \frac{G^2 t}{2}, \text{ as } k \to \infty,$$

and for diminishing step size, we have:

$$\sum_{i=0}^{k} t_i^2 \leq 0, \sum_{i=0}^{k} t_i = \infty$$

therefore,

$$\frac{R^2 + G^2 \sum_{i=0}^{k} t_i^2}{2 \sum_{i=0}^{k} t_i} \to 0, \text{ as } k \to \infty,$$

■

So, consider taking $t_i = R/(G\sqrt{k})$, for all $i = 1, ..., k$. Then we can obtain the following bound:

$$\frac{R^2 + G^2 \sum_{i=0}^{k} t_i^2}{2 \sum_{i=0}^{k} t_i} = \frac{RG}{\sqrt{k}}.$$

That is, subgradient method has convergence rate of $O(1/\sqrt{k})$, and to get $f(x_{best}^{(k)}) - f(x^*) \leq \epsilon$, needs $O(1/\epsilon^2)$ iterations.