CS292A Convex Optimization: Gradient Methods and Online Learning **Spring 2019**

Lecture 4: April 18

Lecturer: Yu-Xiang Wang *Scribes: Yuqing Zhu*

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.

4.1 Recap of Gradient Descent

Consider the following minimization problem:

$$\min_x f(x) \tag{4.1}$$

Gradient descent can be used to minimize $f(x)$ iterating the following steps:

$$x^{(k)} = x^{(k-1)} - t_k \cdot \nabla f(x^{(k-1)}), k = 1, 2, 3... \tag{4.2}$$

4.1.1 Comparison between Different Condition

Strongly convex and smooth is a strong condition. Polyak- Lojasiewicz(PL) condition is overlapping with convex condition, it requires the gradient rate to be big and is still ratively restrict. if f is convex and smooth, then $(RSC) = (PL) = (QC) = errorbound(EB)$. For general smooth functions, we have $(RSI) \to (EB) = (PL) \to (QG)$. Gradient descent(GD) is not optimal, when know the conditonal number L, m, we could use Acceleration GD, which will have a square root improvement, and is optimal for first order method. Question: Why we learn GD instead of AGD? AGD is a gradient descent method, it doen't guarantee grdient descent in every iteration. Up to now we talk about are convex and smooth, we are working with function that are not smooth.

4.2 Subgradients

Definition 4.1 *A subgradient of a convex function f at x is any $g \in \mathcal{R}^n$ such that*

$$f(y) \geq f(x) + g^T(y-x) for\ all y$$

- For convex functions, such g always exists (Subgradients need to be on the ralative interior of dom$f$)

$$f(x) = \begin{cases} -(1-||x||^2)^{\frac{1}{2}} & \text{if } ||x||_2 \leq 1; \\ \infty & \text{otherwise;} \end{cases}$$

- If f is differentialble at x, then f has a unique subgradient at x which is exactly $\nabla f(x)$

- although the same definition of subgradients can also work for nonconvex functions, subgradients may not exist at certain locations, even if they may be smooth.
- Two examples of nonconvex with no subgradients everywhere: $f(x) = -x^2$ and $f(x) = x^3$.

Example 4.2 *where subgradients not exists. e.x. a concave function, or $f(x) = x^q$, where $q > 1$ is a constant, for arbitrary line which is tangent to $f(x)$, will always cross $f(x)$.*

4.2.1 Examples of Subgradients

- **Absolute value** Consider $f(x) = |x|$, for $x \neq 0$, unique subgradient $g = \text{sign}(x)$, otherwise subgradient g is any element between $[-1, 1]$.
- l_2 **norm** Consider $f(x) = ||x||_2$, for $x \neq 0$, unique subgradient $g = x/||x||_2$, otherwise subgradient is any element of $\{z : ||z||_2 \leq 1\}$.
- l_1 **norm** Consider $f(x) = ||x||_1$,for $x_i \neq 0$, unique subgradient $g_i = \text{sign}(x)$, otherwise subgradient g_i is any element between $[-1, 1]$.
- **Pointwise max of two differentiable convex functions** The function has the form $f_1, f_2 : \mathbb{R}^n -> \mathbb{R}, f(x) = \max\{f_1(x), f_2(x)\}$.The function is differentiable at any location where $f_1(x) > f_2(x) or f_1(x) < f_2(x)$. At these locations the subgradient is uniquely equal to the gradient of the larger function. However at locations where $f_1(x) = f_2(x)$, the function becomes nondifferentialble.

$$g(x) = \begin{cases} \nabla f_1(x) & \text{if} f_1(x) \geq f_2(x) \\ \nabla f_2(x) & \text{if} f_1(x) \leq f_2(x) \\ t\nabla f_1(x) + (1-t)\nabla f_2(x), t \in [0, 1] & \text{if} f_1(x) = f_2(x) \end{cases}$$

4.3 Subdifferential

Set of all subgradients of convex f is called the subdifferential,

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

4.3.1 Property of Subdifferential

- It's nonempty if f is convex.
- $\partial f(x)$ is closed and convex even for nonconvex f.
- If f is differentialble at x, then $\partial f(x) = \{\nabla f(x)\}$.
- if $\partial f(x) = \{g\}$, then f is differentiable at x and $\nabla f(x) = g$.

4.3.2 Connection of Convexity Geometry

Given a convex set $C \subseteq \mathcal{R}^n$, consider indicator function $I_C : \mathcal{R}^n -> \mathcal{R}$, where

$$I_C(x) = \begin{cases} I\{x \in C\} = 0 & \text{if } x \in C \\ \infty & \text{otherwise} \end{cases}$$

Proof: By the definition of subgradient g, $I_C(y) \geq I_C(x) + g^T(y-x)$, since for $y \notin C, I_c(y) = \infty$, so we have $0 \geq g^T(y-x), \forall y \in C$. ■

4.3.3 Subgradient Calcus

- Scaling: $\partial(af) = a\dot{\partial} f$, provided $a \geq 0$(if $a \leq 0$, it will turn the function into a concave function).
- Addition: $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$.
- Affine composition: if $f(x) = g(Ax+b)$, then $\partial g = A^T \partial f(Ax+b)$.
- Finite pointwise maximum: if $f(x) = \max_{i \in [1,m]} f_i(x)$, then

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

- Norm : $f(x) = ||x||_p$, Let q be such that $1/p + 1/q = 1$, then $||x||_p = \max_{||z||_q \leq 1} z^T x$, which is also the definition of dual norm. Then we have

$$\partial f(x) = argmax_{||z||_q \leq 1} z^T x. \text{(This is called a polar operator from Yaolin Yu's NIPS'13)}$$

4.3.4 Importance of Subgradient

- For convex analysis, optimality characterization via subgradients. That is for any f(convex or not),

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

  This is called the subgradient optimality condition. Since

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

- For convex optimization, if you can compute subgradients, then you can minimize any convex function.

4.3.5 Derivation of First-order Optimality

Consider a constrained minimization problem:

$$\min_x f(x) + I_C(x)$$

By apply subgradient optimality we have $0 \in \partial(f(x) + I_C(x))$.

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

Note: the condition $0 \in \partial(f(x) + I_C(x)$ is a fully general condition for optimality in convex problems. But it's not always easy to work with (KKT conditions, later, are easier)

4.3.6 Example: Lasso Optimality Conditions

Given $y \in \mathbb{R}^n, x \in \mathbb{R}^{n\times p}$, lasso problem is

$$\min_{\beta} \frac{1}{2}||y - X\beta||_2^2 + \lambda||\beta||_1, \text{where } \lambda \geq 0$$

Follow subgradient optimality, we have

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

4.3.7 Example: Soft-thresholding

For a simplified lasso problem, the sulution is $\beta = S_\lambda(y)$,where S_λ is the soft-thresholding operator.

4.3.8 Example: Distance to a Convex Set

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

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

Write $dist(x, C) = ||x-P_C(x)||_2$, where $P_C(x)$ is the projection of x onto C. It turns out when $dist(x, C) > 0$,

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

Proof: Suppose $u = P_C(x)$, then by first-order optimality conditions for a projection, we have

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

Hence $C \subset H = \{y : (x - u)^T(y - u) \leq 0\}$. Then we have

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

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

example.png

Figure 4.1: Illustration of the subgradients of three example nonsmooth functions. From left to right: 1. absolute value; 2. l_2 norm; 3. l_1 norm. 4. pointwise max of two differentiable convex functions.