CS 7545: Machine Learning Theory **Fall 2018**

Lecture 2: Convex Analysis

Lecturer: Jacob Abernethy *Scribes: Haodong Sun, Kun Zheng*

Main Content

- Review Norm and Inequality
- Lipschitz function
- Convexity (strong convexity, smoothness)
- Bregman divergence
- Fenchel conjugate

2.1 Review Norm and Inequality

- **Definition 2.1 (norm)** $\|\cdot\|$ *is a* ***norm*** *on* $\mathbb{R}^n$ *if*
  - $\|\vec{x}\| \geq 0 \quad \forall \vec{x} \in \mathbb{R}^n$
  - $\|\vec{x}\| = 0 \iff \vec{x} = 0$
  - $\|\vec{x} + \vec{y}\| \leq \|\vec{x}\| + \|\vec{y}\|$ ***(triangular inequality)***
- **Definition 2.2 (dual norm)** *The* ***dual norm*** $\|\cdot\|_\star$ *on* $\mathbb{R}^n$ *is*

$$\|\vec{y}\|_\star = \sup_{\vec{x}:\|\vec{x}\|=1} \vec{x}^\top \vec{y}$$

(Exercise)

- $\|\vec{x}\| := \|\vec{x}\|_2 = (\sum_i x_i^2)^{1/2} \Rightarrow \|\vec{y}\|_\star = \|\vec{y}\|_2$
- $\|\vec{x}\| := \|\vec{x}\|_p = (\sum_i x_i^p)^{1/p} \Rightarrow \|\vec{y}\|_\star = \|\vec{y}\|_q$ where $\frac{1}{p} + \frac{1}{q} = 1$

- **Theorem 2.3 (Hölder's inequality)** *For any* $\vec{x}, \vec{y} \in \mathbb{R}^n$*, any norm* $\|\cdot\|$*,* $\|\vec{x}^\top \vec{y}\|_1 \leq \|\vec{x}\|\|\vec{y}\|_\star$

  Proof: If $\vec{x} = 0$, it's trivial. Assume that $\vec{x} \neq 0$: Let $\vec{w} = \frac{\vec{x}}{\|\vec{x}\|}, \|\vec{w}\| = 1$.

  By definition of dual norm,

$$\|\vec{y}\|_\star \geq \vec{w}^\top \vec{y} = \frac{\vec{x}^\top \vec{y}}{\|\vec{x}\|}.$$

■

- **Theorem 2.4 (Cauchy-Schwarz inequality)**

$$\|\vec{x}^\top \vec{y}\|_1 \leq \|\vec{x}\|_2\|\vec{y}\|_2, \quad i.e. |\sum_i x_i y_i| \leq (\sum_i x_i^2)^{\frac{1}{2}}(\sum_i y_i^2)^{\frac{1}{2}}$$

  Proof:

1. Apply **Hölder's inequality** with $\|\cdot\| := \|\cdot\|_\star = \|\cdot\|_2$.
2. Direct proof:

$$(\sum_i x_i^2)(\sum_i y_i^2) - (\sum_i x_i y_i)^2 = \frac{1}{2}\sum_{i,j}(x_i y_j - x_j y_i)^2 \geq 0$$

■

2.2 Lipchitz Function

- **Definition 2.5 (gradient)** *Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be a differentiable function. The* ***gradient*** *of f at $\vec{x} \in \mathbb{R}^n$ is the vector of partial derivatives:*

$$\nabla f(\vec{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(\vec{x}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\vec{x}) \end{bmatrix} \in \mathbb{R}^n$$

e.g.

$$\begin{aligned} f(\vec{x}) &= \frac{1}{2}\vec{x}^\top M\vec{x} \quad (M = M^\star) \\ \nabla f(\vec{x}) &= M\vec{x} \\ \nabla^2 f(\vec{x}) &= M \end{aligned}$$

- **Definition 2.6 (Hessian)** *The* ***Hessian*** *of f at $\vec{x} \in \mathbb{R}^n$ is the symmetric matrix of second partial derivatives:*

$$\nabla^2 f(\vec{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1^2}(\vec{x}) & \cdots & \frac{\partial f}{\partial x_1 \partial x_n}(\vec{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_n \partial x_1}(\vec{x}) & \cdots & \frac{\partial f}{\partial x_n^2}(\vec{x}) \end{bmatrix}$$

where $\frac{\partial f}{\partial x_i \partial x_j} = \frac{\partial f}{\partial x_j \partial x_i}$.

- **Definition 2.7 (L-Lipschitz)** *A function $f : \mathbb{R}^n \to \mathbb{R}$ is* ***L-Lipschitz*** *(for some $L > 0$) with respect to $\|\cdot\|$ if*

$$|f(\vec{x}) - f(\vec{y})| \leq L\|\vec{x} - \vec{y}\|, \quad \forall \vec{x}, \vec{y} \in \mathbb{R}^n$$

(f grows at most linearly)

- **Theorem 2.8**

If f is differentiable: f is ***L-Lipschitz*** $\iff \|\nabla f(\vec{x})\|_\star \leq L, \quad \forall \vec{x} \in \mathbb{R}^n$

Proof:

"$\Rightarrow$" Suppose f is **L-Lipschitz**. The directional derivative of f at $\vec{x}^\star$ along direction $\vec{u}$ is

$$\begin{aligned} \nabla f(\vec{x})^\top \vec{u} &= \lim_{h\to 0} \frac{f(\vec{x} + h\vec{u}) - f(\vec{x})}{h} \\ &\leq \lim_{h\to 0} \frac{L\|\vec{x} + h\vec{u} - \vec{x}\|}{h} \\ &= L\|\vec{u}\| \end{aligned}$$

By the definition of dual norm, we have $\|\nabla f(\vec{x})\|_\star \leq L$.

"$\Leftarrow$" Suppose $\|\nabla f(\vec{x})\|_\star \leq L$. For any x, y, by the **mean value theorem**, there exists $t \in [0,1]$, s.t.

$$f(\vec{x}) = f(\vec{y}) + \nabla f(t\vec{x} + (1-t)\vec{y})^\top (\vec{x} - \vec{y})$$

then by Hölder's inequality we have

$$\begin{aligned}|f(\vec{x}) - f(\vec{y})| &= \|\nabla f(t\vec{x} + (1-t)\vec{y})^\top (\vec{x} - \vec{y})\| \\ &\leq \|\nabla f(t\vec{x} + (1-t)\vec{y})\|_\star \|\vec{x} - \vec{y}\| \\ &\leq L\|\vec{x} - \vec{y}\|\end{aligned}$$

■

2.3 Convexity

In this section, the lecturer introduced firstly the concept of convex set and three ways to define a convex function. Following the definition of convex function are definition of strongly convex and strongly smooth, accompanied with a simple example of quadratic function.

The definition of convex set and convex function come at first:

- **Definition 2.9 (convex set)** *A set $U \in R^n$ is **convex** if for any $\vec{x}, \vec{y} \in U$ and any $t \in [0,1]$,*

$$t\vec{x} + (1-t)\vec{y} \in U.$$

- **Definition 2.10 (convex function)** *A function $f : \vec{x} \to R$ is **convex function** if for any $\vec{x}, \vec{y}$ and any $t \in [0,1]$,*

$$f(t\vec{x} + (1-t)\vec{y}) \leq tf(\vec{x}) + (1-t)f(\vec{y})$$

- **Definition 2.11 (convex function)** *If f is differentiable, then f is **convex function** if and only if for any $\vec{x}, \vec{y} \in \mathrm{dom}(f)$,*

$$f(\vec{x}) \geq f(\vec{y}) + \nabla f(\vec{y})^\top (\vec{x} - \vec{y})$$

- **Definition 2.12 (convex function)** *If f is twice-differentiable, then f is **convex function** if and only if for any $\vec{x} \in \mathrm{dom}(f)$,*

$$\nabla^2 f(\vec{x}) \geq 0$$

The following are some properties of convex function

- If f is convex, then it satisfies the Jensen's inequality: For any random variable $\vec{x} \in \mathrm{dom} f, E[f(\vec{x})] \geq f(E[\vec{x}])$.
- If $g(\vec{x}, \vec{y})$ is convex in $\vec{x}$ (for example: $g(\vec{x}, \vec{y}) = ||\vec{x}||^2 - ||\vec{y}||^2$), then $f_1(\vec{x}) = E_{\vec{y}}[g(\vec{x}, \vec{y})]$ and $f_2(\vec{x}) = \sup_{\vec{y}} g(\vec{x}, \vec{y})$ are both convex.

Here comes the definition of strongly convex and strongly smooth.

- **Definition 2.13 (strongly convex)** *A differentiable function f is m-**strongly convex** if there exists $m > 0$ such that for any $\vec{x}, \vec{y} \in \text{dom} f$,*

$$f(\vec{x}) \geq f(\vec{y}) + \nabla f(\vec{y})^\top (\vec{x} - \vec{y}) + \frac{m}{2}||\vec{x} - \vec{y}||^2.$$

  *If f is also twice differentiable, then f is m-**strongly convex** if and only if $\nabla^2 f(\vec{x}) \geq mI$, where I is the identity matrix.*
  Intuition: f always has a quadratic lower bound and the function grows at least quadratically.

- **Definition 2.14 (strongly smooth)** *A differentiable function f is l-**strongly smooth** if there exists $l > 0$ such that for any $\vec{x}, \vec{y} \in \text{dom} f$,*

$$f(\vec{x}) \leq f(\vec{y}) + \nabla f(\vec{y})^\top (\vec{x} - \vec{y}) + \frac{l}{2}||\vec{x} - \vec{y}||^2.$$

  *If f is also twice differentiable, then f is l-**strongly smooth** if and only if $\nabla^2 f(\vec{x}) \leq lI$, where I is the identity matrix. Intuition: f always has a quadratic upper bound and the function grows at most quadratically.*

- Example 1: Here is a special example of quadratic function $f(\vec{x}) = \frac{1}{2}||\vec{x}||^2$. Because we have $\nabla^2 f(\vec{x}) = I$, the function is both 1-strongly convex and $1-$strongly smooth.

- Example 2: Here is another example of $f(\vec{x}) = \frac{1}{2}\vec{x}^\top M\vec{x}$, which has second derivative as $\nabla^2 f(\vec{x}) = M$. Since $\lambda_{min}(M)I \leq M \leq \lambda_{max}(M)I$, the function is $\lambda_{max}(M)$-strongly smooth and $\lambda_{min}(M)$-strongly convex.

2.4 Bregman Divergence

The section gives an introduction to Bregman divergence. In the class, several properties of Bregman divergence were also discussed accompanied with examples as is needed.

First comes the definition of Bregman divergence:

- **Definition 2.15 (Bregman divergence)** *Let $f : R^n \to R$ be a differentiable function, then the **Bregman divergence** is a mapping function $D_f : R^n \times R^n \to R$ given by,*

$$D_f(\vec{x}, \vec{y}) = f(\vec{x}) - f(\vec{y}) - \nabla f(\vec{y})^\top (\vec{x} - \vec{y}).$$

Then comes 3 example of Bregman divergence:

- Example 1: When $f(\vec{x}) = \frac{1}{2}||\vec{x}||_2^2$, then $D_f(\vec{x}, \vec{y}) = \frac{1}{2}||\vec{x} - \vec{y}||_2^2$.

- Example 2: When $f(\vec{x}) = \frac{1}{2}\vec{x}^\top M\vec{x}, D_f(\vec{x}, \vec{y}) = \frac{1}{2}(\vec{x} - \vec{y})^\top M(\vec{x} - \vec{y})$.

- Example 3 (**Kullbeck-Lesiblez divergence**): When $f(\vec{p}) = \sum_{i=1}^n p_i \log p_i, D_f(p, q) = \sum_{i=1}^n p_i \log \frac{p_i}{q_i}$. In this case, the Bregman divergence here is also called **Kullbeck-Lesiblez divergence**, or **K-L divergence**.

The following are some important properties of Bregman divergence:

- Property 1: If f is convex, then for any $\vec{x}, \vec{y} \in \text{dom} f, D_f(\vec{x}, \vec{y}) \geq 0$.

- Property 2: for any $\vec{x} \in \text{dom} f, D_f(\vec{x}, \vec{x}\}) = 0$.

- Property 3: In general, $D_f(\vec{x}, \vec{y}) \neq D_f(\vec{y}, \vec{x})$.

- Property 4: If f is m-strongly convex, then $D_f(\vec{x}, \vec{y}) \geq \frac{m}{2}||\vec{x} - \vec{y}||^2$ for any $\vec{x}, \vec{y} \in \text{dom}f$.
- Property 5: If f is l-strongly smooth, then $D_f(\vec{x}, \vec{y}) \leq \frac{l}{2}||\vec{x} - \vec{y}||^2$ for any $\vec{x}, \vec{y} \in \text{dom}f$.
- Property 6: If f is m-strongly convex, then $||\nabla f(\vec{x}) - \nabla f(\vec{y})|| \geq m||\vec{x} - \vec{y}||$.
- Property 7: If f is convex and l-strongly smooth, then $||\nabla f(\vec{x}) - \nabla f(\vec{y})|| \leq l||\vec{x} - \vec{y}||$. That is to say, ∇f is l-Lipschitz continuous.