CS 7545: Machine Learning Theory **Fall 2019**

Lecture 2: Convex Analysis

Lecturer: Jacob Abernethy *Scribes: Cyrille Combettes and Gabriel Nakajima An*

2.1 Vector analysis

Definition 2.1 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function and $\vec{x} \in \mathbb{R}^n$. The **gradient of** f **at** $\vec{x}$ is*

$$\nabla f(\vec{x}) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(\vec{x}) & \dots & \frac{\partial f}{\partial x_n}(\vec{x}) \end{pmatrix}.$$

Remark 2.2 *The gradient is correctly represented as a row vector but during analysis and proofs, we may represent it as a column vector to make it clear that the products are matrix vector products.*

Definition 2.3 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function and $\vec{x} \in \mathbb{R}^n$. The **Hessian of** f **at** $\vec{x}$ is*

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

Fact 2.4 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function and $\vec{x} \in \mathbb{R}^n$. Then $\nabla^2 f(\vec{x})$ is a symmetric matrix, because for all $i, j \in [\![1, n]\!]$, it holds*[1]

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\vec{x}) = \frac{\partial^2 f}{\partial x_j \partial x_i}(\vec{x}).$$

2.2 Convexity

Definition 2.5 *Let $\mathcal{U} \subseteq \mathbb{R}^n$. The set $\mathcal{U}$ is a **convex** set if*

$$\forall \vec{x}, \vec{y} \in \mathcal{U}, \forall t \in [0,1] : \quad t\vec{x} + (1-t)\vec{y} \in \mathcal{U},$$

i.e., any line segment connecting two points in $\mathcal{U}$ is in $\mathcal{U}$.

Definition 2.6 *Let $\mathcal{U} \subseteq \mathbb{R}^n$ be a convex set and $f : \mathcal{U} \to \mathbb{R}$. f is **convex** if*

$$\forall \vec{x}, \vec{y} \in \mathcal{U}, \forall t \in [0,1] : \quad f((1-t)\vec{x} + t\vec{y}) \leq (1-t)f(\vec{x}) + tf(\vec{y}),$$

i.e., any line segment connecting two points on the graph of f is above the graph of f.

Claim 2.7 *Let $\mathcal{U} \subseteq \mathbb{R}^n$ be a convex set and $f : \mathcal{U} \to \mathbb{R}$ be a differentiable function. Then f is convex if and only if*

$$\forall \vec{x}, \vec{y} \in \mathcal{U} : \quad f(\vec{x}) \geq f(\vec{y}) + \langle \nabla f(\vec{y}), \vec{x} - \vec{y} \rangle,$$

i.e., f is always above its tangents.

[1] See Schwarz' theorem.

Proof: Exercise 1. ■

Notation 2.8 *Let $M \in \mathbb{R}^{n\times n}$ be a square matrix. We denote M* ***is positive semi-definite*** *by $M \succeq 0$ and M* ***is positive definite*** *by $M \succ 0$. We denote $M \preceq 0$ if $-M \succeq 0$ and $M \prec 0$ if $-M \succ 0$.*

Claim 2.9 *Let $\mathcal{U} \subseteq \mathbb{R}^n$ be a convex set and $f : \mathcal{U} \to \mathbb{R}$ be a twice differentiable function. Then f is convex if and only if*

$$\forall \vec{x} \in \mathcal{U}, \nabla^2 f(\vec{x}) \succeq 0.$$

Remark 2.10 *In order to check if a twice differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is convex or not at a point $\vec{x} \in \mathbb{R}^n$, it is sometimes easier to check Claim 2.9 than Claim 2.7, because it only requires studying one object: the Hessian matrix $\nabla^2 f(\vec{x})$ (although computing its eigenvalues may be costly).*

Definition 2.11 *Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}$, and $c \geq 0$. f is* ***c-Lipschitz with respect to*** *$\|\cdot\|$ if*

$$\forall \vec{x}, \vec{y} \in \mathbb{R}^n : \quad |f(\vec{x}) - f(\vec{y})| \leq c\|\vec{x} - \vec{y}\|. \tag{2.1}$$

A Lipschitz function grows at most and at least linearly.

Claim 2.12 *A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is c-Lipschitz with respect to $\|\cdot\|$ if and only if*

$$\forall \vec{x} \in \mathbb{R}^n : \quad \|\nabla f(\vec{x})\|_* \leq c$$

where $\|\cdot\|_*$ is the dual norm associated with $\|\cdot\|$.

Proof: $\boxed{\Rightarrow}$ Suppose that f is c-Lipschitz with respect to $\|\cdot\|$ and let $\vec{x}, \vec{u} \in \mathbb{R}^n$. The quantity $\langle \nabla f(\vec{x}), \vec{u} \rangle$ is the directional derivative of f at $\vec{x}$ in the direction of $\vec{u}$. We have

$$\begin{aligned}
\langle \nabla f(\vec{x}), \vec{u} \rangle &= \lim_{\substack{h \to 0 \\ h > 0}} \frac{f(\vec{x} + h\vec{u}) - f(\vec{x})}{h} \\
&\leq \lim_{\substack{h \to 0 \\ h > 0}} \frac{c\|\vec{x} + h\vec{u} - \vec{x}\|}{h} \qquad \text{by Equation (2.1)} \\
&= c \lim_{\substack{h \to 0 \\ h > 0}} \frac{h\|\vec{u}\|}{h} = c\|\vec{u}\|
\end{aligned}$$

thus

$$\|\nabla f(\vec{x})\|_* = \sup_{\|u\| \leq 1} \langle \nabla f(\vec{x}), \vec{u} \rangle \leq \sup_{\|u\| \leq 1} c\|\vec{u}\| = c.$$

Therefore,

$$\forall \vec{x} \in \mathbb{R}^n : \|\nabla f(\vec{x})\|_* \leq c.$$

$\boxed{\Leftarrow}$ Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$ and $c \geq 0$, and suppose that $\|\nabla f(\vec{x})\|_* \leq c$ for all $\vec{x} \in \mathbb{R}^n$. By the mean value theorem, there exists $t \in [0, 1]$ such that

$$f(\vec{x}) = f(\vec{y}) + \langle \nabla f(\vec{z}), \vec{x} - \vec{y} \rangle,$$

where $z = (1 - t)\vec{x} + t\vec{y}$. By Hölder's inequality, we conclude that

$$\begin{aligned}
|f(\vec{x}) - f(\vec{y})| &= |\langle \nabla f(\vec{z}), \vec{x} - \vec{y} \rangle| \\
&\leq \|\nabla f(\vec{z})\|_* \|\vec{x} - \vec{y}\| \\
&\leq c\|\vec{x} - \vec{y}\|.
\end{aligned}$$

Therefore, f is c-Lipschitz with respect to $\|\cdot\|$. ■

Theorem 2.13 (Jensen's inequality) *Let $\mathcal{U} \subseteq \mathbb{R}^n$ be a convex set, $f : \mathcal{U} \to \mathbb{R}$ be a convex function, and X be a random variable on $\mathcal{U}$. Then*

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$

Properties 2.14

1. *f convex, g convex $\Rightarrow$ $f + g$ convex.*
2. *$\alpha \geq 0$, f convex $\Rightarrow$ αf convex.*
3. *f convex, g convex $\Rightarrow$ $h := \max\{f, g\}$ convex.*
4. *$g(\vec{x}, \vec{y})$ jointly convex in $\vec{x}, \vec{y}$ $\Rightarrow$ $f(\vec{x}) := \inf_{\vec{y}} g(\vec{x}, \vec{y})$ is convex.*

Definition 2.15 *Let $\mathcal{U} \subseteq \mathbb{R}^n$ be a convex set and $f : \mathcal{U} \to \mathbb{R}$. f is **concave** if $-f$ is convex.*

Example 2.16 log *is a concave function.*

Theorem 2.17 (Young's inequality) *Let $p, q > 1$ be such that $\frac{1}{p} + \frac{1}{q} = 1$. Then*

$$\forall a, b > 0 : \quad ab \leq \frac{a^p}{p} + \frac{b^q}{q}.$$

Proof: Let $a, b > 0$. By Jensen's inequality applied to the convex function $-\log$,

$$\begin{aligned} \log(ab) &= \log(a) + \log(b) \\ &= \frac{p}{p}\log(a) + \frac{q}{q}\log(b) \\ &= \frac{1}{p}\log(a^p) + \frac{1}{q}\log(b^q) \\ &\leq \log\left(\frac{1}{p}a^p + \frac{1}{q}b^q\right). \end{aligned}$$

Therefore, by applying the monotonically increasing function exp to both sides, we obtain

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}.$$

■

Definition 2.18 *Let $f : \mathrm{dom}(f) \to \mathbb{R}$ be a differentiable function and $\alpha > 0$. f is α-**strongly convex** if*

$$\forall \vec{x}, \vec{y} \in \mathrm{dom}(f) : \quad f(\vec{x}) \geq f(\vec{y}) + \langle \nabla f(\vec{y}), \vec{x} - \vec{y} \rangle + \frac{\alpha}{2}\|\vec{x} - \vec{y}\|^2.$$

A strongly convex function grows at least quadratically.

Claim 2.19 *Let $f : \mathrm{dom}(f) \to \mathbb{R}$ be a twice differentiable function. Then f is α-strongly convex if and only if*

$$\forall \vec{x} \in \mathrm{dom}(f) : \quad \nabla^2 f(\vec{x}) - \alpha I \succeq 0.$$

Definition 2.20 *Let $f : \mathrm{dom}(f) \to \mathbb{R}$ be a differentiable function and $\alpha > 0$. f is α-**smooth** if*

$$\forall \vec{x}, \vec{y} \in \mathrm{dom}(f) : \quad f(\vec{x}) \leq f(\vec{y}) + \langle \nabla f(\vec{y}), \vec{x} - \vec{y} \rangle + \frac{\alpha}{2}\|\vec{x} - \vec{y}\|^2.$$

A smooth function grows at most quadratically.

Claim 2.21 *Let $f : \mathrm{dom}(f) \to \mathbb{R}$ be a twice differentiable function. Then f is α-smooth if and only if*

$$\forall \vec{x} \in \mathrm{dom}(f): \quad \nabla^2 f(\vec{x}) - \alpha I \preceq 0.$$

Example 2.22 *Let $M \in \mathbb{R}^{n\times n}$ be a symmetric positive definite matrix and $f(\vec{x}) = \frac{1}{2}\vec{x}^\top M\vec{x}$. Denote $\lambda_{\min}$ and $\lambda_{\max}$ the smallest and largest eigenvalues of M respectively. Then f is $\lambda_{\min}$-strongly convex and $\lambda_{\max}$-smooth. Note that $\nabla f(\vec{x}) = \vec{x}^\top M$ (see Remark 2.2) and $\nabla^2 f(\vec{x}) = M$ for all $\vec{x} \in \mathbb{R}^n$.*

2.3 Bregman divergence

Definition 2.23 *Let $f : \mathcal{U} \to \mathbb{R}$ be a convex differentiable function and $\vec{x}, \vec{y} \in \mathcal{U}$. The **Bregman divergence of f from $\vec{x}$ to $\vec{y}$** is*

$$D_f(\vec{x}, \vec{y}) := f(\vec{x}) - f(\vec{y}) - \langle \nabla f(\vec{y}), \vec{x} - \vec{y}\rangle.$$

It measures the distance at $\vec{x}$ between the graph of f and its tangent at $\vec{y}$, i.e. the distance between $f(\vec{x})$ and $f(\vec{y}) + \langle \nabla f(\vec{y}), \vec{x} - \vec{y}\rangle$.

Example 2.24

1. *Let $f(\vec{x}) = \frac{1}{2}\|\vec{x}\|_2^2$. Then*

$$\forall \vec{x}, \vec{y} \in \mathbb{R}^n: \quad D_f(\vec{x}, \vec{y}) = \frac{1}{2}\|\vec{x} - \vec{y}\|_2^2.$$

   Note that this is the only situation where the Bregman divergence is quadratic (see Fact 2.25).

2. *Let $\Delta^n = \{\vec{p} \in \mathbb{R}^n \mid \sum_{i=1}^n p_i = 1, p_i \geq 0\}$ denote the unit simplex in dimension n and $f(\vec{p}) = \sum_{i=1}^n p_i \log p_i$ denote the entropy function, with the convention $0 \log 0 = 0$. Then*

$$\forall \vec{p}, \vec{q} \in \Delta^n: \quad D_f(\vec{p}, \vec{q}) = \sum_{i=1}^n p_i \log \frac{p_i}{q_i} =: \mathrm{KL}(\vec{p} \,\|\, \vec{q}),$$

   where KL is the Kullback-Leibler divergence.

Fact 2.25 *Let $f : \mathrm{dom}(f) \subseteq \mathbb{R}^n \to \mathbb{R}$ be a convex differentiable function. Then*

$$\Big(\forall \vec{x}, \vec{y} \in \mathrm{dom}(f): \ D_f(\vec{x}, \vec{y}) = D_f(\vec{y}, \vec{x})\Big) \Leftrightarrow \Big(f \text{ is quadratic}\Big).$$