CS 7545

Machine Learning Theory

Professor Jacob Abernethy

Organized by Yiwen Chen

Contents

Preface

This is the organized course material of CS 7545 Machine Learning Theory from Georgia Institute of Technology. The material is based on the course in Fall 2019. The instructor of this course is professor Jacob Abernethy and the course website can be found via this link.

This course contains many theoretical aspects of machine learning, including decision making, online learning, boosting, regret minimization, etc. Students will get familiar with the frontiers of theoretical machine learning research after finishing this course. It is a totally theoretical course and no programming will be covered.

CS 7545 is a graduate level course, thus it has several prerequisites. Familiarity with Algorithm analysis, linear algebra, probability and statistics is a must to understand the contents of this course. Convex analysis background ~~is strongly recommended but not required~~ is also very important.

Before studying any algorithm and theory, we will first review some math aspects and point out some notations that we will use in this course material. Afterwards, we will talk about convex analysis and deviation bound. Online learning algorithms will be covered next. Finally, we will talk about statistical learning theory.

Thanks to professor Jacob Abernethy for his instruction.

Chapter 1

Math Review

In this chapter, we will briefly review several aspects of linear algebra, norm, etc. The following notations will be used throughout the course.

- M: Matrix (size $\mathbb{R}^{m \times n}$).
- M^T: Transpose Matrix of M.
- $\mathbf{x}$: Vector (size $\mathbb{R}^n$).
- $\mathbf{x}_i$: i-th element of vector $\mathbf{x}$ (size $\mathbb{R}$).

1.1 Positive Semidefinite and Positive Definite Matrices

We will first define Positive semidefinite (PSD) and positive definite matrices.

Definition 1.1 (**Positive Semidefinite Matrix).** *A matrix $M \in \mathbb{R}^{n \times n}$ is said to be positive semidefinite (PSD) if it satisfies the following two conditions.*

1. *M is a symmetric matrix. That is $M^T = M$.*
2. *For all $\mathbf{x} \in \mathbb{R}^n$, we have that $\mathbf{x}^T M \mathbf{x} \geq 0$. This condition is equivalent to all eigenvalues of M being non-negative.*

We can denote matrix M being positive semidefinite (PSD) as $M \succeq 0$.

Definition 1.2 (**Positive Definite Matrix).** *Similar to above definition, a matrix M is said to be positive definite (PD) if it satisfies the following two conditions:*

1. *M is a symmetric matrix. That is $M^T = M$.*
2. *For all $\mathbf{x} \in \mathbb{R}^n$, we have that $\mathbf{x}^T M \mathbf{x} > 0$. This condition is equivalent to all eigenvalues of M being positive.*

We can denote matrix M being positive definite (PD) as $M \succ 0$.

Notice that the only difference between positive semidefinite matrices and positive definite matrices is whether the eigenvalues could be zero. In addition, all PD matrices are PSD, while PSD matrices may not be PD.

Notation 1.3. *We denote $M \preceq 0$ if $-M \succeq 0$ and $M \prec 0$ if $-M \succ 0$.*

1.2 Norm

Definition 1.4 (Norm). *Norm is defined as a function $||\cdot|| : \mathbb{R}^n \to [0, \infty]$ which satisfies the following conditions:*

1. ***Identity of Indiscernibles**: $||\mathbf{x}|| = 0$ if and only if $\mathbf{x} = 0$.*
2. ***Absolute Homogeneity**: $||\alpha\mathbf{x}|| = |\alpha|||\mathbf{x}||$ for all $\mathbf{x} \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$.*
3. ***Triangle Inequality**: $||\mathbf{x} + \mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}||$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.*

Some famous norm functions can be seen following:

1. $\boldsymbol{\ell}_1$ norm: $||\mathbf{x}||_1 = \sum_{i=1}^n |\mathbf{x}_i|$.
2. $\boldsymbol{\ell}_2$ norm: $||\mathbf{x}||_2 = \sqrt[2]{\sum_{i=1}^n \mathbf{x}_i^2}$.
3. $\boldsymbol{\ell}_\infty$ norm: $||\mathbf{x}||_\infty = \max\limits_{1 \leq i \leq n} |\mathbf{x}_i|$.
4. $\boldsymbol{\ell}_p$ norm: $||\mathbf{x}||_p = \sqrt[p]{\sum_{i=1}^n |\mathbf{x}_i|^p}$
5. M norm (suppose $M \in \mathbb{R}^{n \times n}$): $||\mathbf{x}||_M = \sqrt[2]{\mathbf{x}^T M \mathbf{x}}$.

We can see that $\boldsymbol{\ell}_1$, $\boldsymbol{\ell}_2$ and $\boldsymbol{\ell}_\infty$ are special case of $\boldsymbol{\ell}_p$ norm.

We will next prove that $\boldsymbol{\ell}_\infty$ norm satisfies the three condition in norm definition.

Proof. We will just use the norm definition to finish the proof.

1. **Identity of Indiscernibles**:
   Since $||\mathbf{x}||_\infty = \max\limits_{1 \leq i \leq n} |\mathbf{x}_i|$. $||\mathbf{x}||_\infty = 0$ if and only if $\mathbf{x}_i = 0$ for all $i \in [1, n]$.
2. **Absolute Homogeneity**:
   $||\alpha\mathbf{x}||_\infty = \max\limits_{1 \leq i \leq n} |\alpha\mathbf{x}_i| = |\alpha| \max\limits_{1 \leq i \leq n} |\mathbf{x}_i| = |\alpha|||\mathbf{x}||_\infty$.
3. **Triangle Inequality**:
   $||\mathbf{x} + \mathbf{y}||_\infty = \max\limits_{1 \leq i \leq n} |\mathbf{x}_i + \mathbf{y}_i| \leq \max\limits_{1 \leq i \leq n} (|\mathbf{x}_i| + |\mathbf{y}_i|) \leq \max\limits_{1 \leq i \leq n} |\mathbf{x}_i| + \max\limits_{1 \leq j \leq n} |\mathbf{y}_j| = ||\mathbf{x}||_\infty + ||\mathbf{y}||_\infty$.

□

Definition 1.5 (Dual Norm). *Given a norm $||\cdot||$, we can define its dual norm $||\cdot||_*$ as following:*

$$||\mathbf{y}||_* = \sup_{\mathbf{x} \in \mathbb{R}^n, ||\mathbf{x}|| = 1} \langle \mathbf{x}, \mathbf{y} \rangle \tag{1.1}$$

where $\langle , \rangle$ is the inner product of two vectors.

Claim 1.6. *The ℓ_2 norm's dual norm is ℓ_2 norm itself.*

Proof. For any vector $\mathbf{y} \in \mathbb{R}^n$, we have that

$$||\mathbf{y}||_{2,*} = \sup_{\mathbf{x}\in\mathbb{R}^n, ||\mathbf{x}||_2=1} \langle \mathbf{x}, \mathbf{y} \rangle = \sup_{\mathbf{x}\in\mathbb{R}^n} \langle \frac{\mathbf{x}}{||\mathbf{x}||_2}, \mathbf{y} \rangle = \frac{1}{||\mathbf{y}||_2} \langle \mathbf{y}, \mathbf{y} \rangle = ||\mathbf{y}||_2.$$

Notice that here the inner product is maximized when two vectors ($\frac{\mathbf{x}}{||\mathbf{x}||_2}$ and $\mathbf{y}$) take the same direction. □

Claim 1.7. *The ℓ_p norm is dual to ℓ_q norm if and only if the following equation holds:*

$$\frac{1}{p} + \frac{1}{q} = 1 \tag{1.2}$$

Proof. TBD □

Claim 1.8. *For a PD Matrix M, the dual norm of M norm is M^{-1} norm, where M^{-1} is the inverse matrix of M.*

Proof. TBD □

Theorem 1.9 (Hölders Inequality). *Suppose $||\cdot||$ and $||\cdot||_*$ are a pair of dual norms. Then for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we can have that*

$$\langle \mathbf{x}, \mathbf{y} \rangle \leq ||\mathbf{x}||||\mathbf{y}||_*. \tag{1.3}$$

Proof. By definition,

$$||\mathbf{x}||||\mathbf{y}||_* = ||\mathbf{x}|| \sup_{\mathbf{z}\in\mathbb{R}^n, ||\mathbf{z}||_2=1} \langle \mathbf{z}, \mathbf{y} \rangle \geq ||\mathbf{x}|| \langle \frac{\mathbf{x}}{||\mathbf{x}||}, \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle.$$

Here we use the fact that the supremum over $\mathbf{z}$ must be larger or equal than taking any vector $\mathbf{x}$. □

This theorem will be very useful in the upcoming chapters.

Chapter 2

Convex Analysis

2.1 Convex Set and Convex Function

In this section, we will review some concepts in convex analysis.

First, let us define some basic and useful notations.

Definition 2.1 (**Gradient**). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function and let $\mathbf{x} \in \mathbb{R}^n$. We can defined the gradient of f at $\mathbf{x}$ as*

$$\nabla f(\mathbf{x}) = (\frac{\partial f}{\partial \mathbf{x}_1}(\mathbf{x}), \cdots, \frac{\partial f}{\partial \mathbf{x}_n}(\mathbf{x})). \tag{2.1}$$

Remark 2.2. *Notice that here we define gradient as a row vector ($\mathbb{R}^{1\times n}$). However, for convenience, we may abuse this notation and use it as a column vector ($\mathbb{R}^{n\times 1}$) in the following contents.*

Definition 2.3 (**Hessian**). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function and let $\mathbf{x} \in \mathbb{R}^n$. Then the Hessian of f at $\mathbf{x}$ is defined as*

$$\nabla^2 f(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(\mathbf{x}) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(\mathbf{x}) & \cdots & \frac{\partial^2 f}{\partial x_n^2}(\mathbf{x}). \end{pmatrix} \tag{2.2}$$

Notice that if f is a twice differentiable function and $\mathbf{x} \in \mathbb{R}^n$, then $\nabla^2 f(\mathbf{x})$ is a symmetric matrix since for all $i, j \in [1, n]$, it has

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

Next, we will talk about convexity, its definition and several useful facts.

Definition 2.4 (**Convex Set**). *Let $\mathcal{K} \subset \mathbb{R}^n$, set $\mathcal{K}$ is called a convex set if $\forall \mathbf{x}, \mathbf{y} \in \mathcal{K}$ and $\forall t \in [0, 1]$, we have*

$$t\mathbf{x} + (1 - t)\mathbf{y} \in \mathcal{K}. \tag{2.3}$$

The meaning of this definition is that for any two arbitrary points in the set $\mathcal{K}$, the straight line between these two points are all contained in $\mathcal{K}$.

Definition 2.5 (Convex Function). *Let $\mathcal{K} \subset \mathbb{R}^n$ be a convex set and $f : \mathcal{K} \to \mathbb{R}^n$ be a differentiable function. Then function f is convex if $\forall \mathbf{x}, \mathbf{y} \in \mathcal{K}$ and $\forall t \in [0, 1]$, we have that*

$$f((1-t)\mathbf{x} + t\mathbf{y}) \leq (1-t)f(\mathbf{x}) + tf(\mathbf{y}). \tag{2.4}$$

Claim 2.6. *Let $\mathcal{K} \subset \mathbb{R}^n$ be a convex set and $f : \mathcal{K} \to \mathbb{R}^n$ be a differentiable function. Then function f is convex if and only if $\forall \mathbf{x}, \mathbf{y} \in \mathcal{K}$, we have that*

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \langle f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle. \tag{2.5}$$

An intuition behind this claim is that f is convex if and only if the first order approximation of f at any point $\mathbf{y}$ is not larger than the function itself.

Claim 2.7. *Let $\mathcal{K} \subset \mathbb{R}^n$ be a convex set and $f : \mathcal{K} \to \mathbb{R}$ be a twice differentiable function. Then f is convex if and only if $\forall \mathbf{x} \in \mathcal{K}$,*

$$\nabla^2 f(\mathbf{x}) \succeq 0. \tag{2.6}$$

Remark 2.8. *It can be seen that above three formulas (2.4), (2.5) and (2.6) are equivalent. However, in reality, if f is twice differentiable, using (2.6) is much simpler than using (2.4) or (2.5), because it only contains one variable in $\mathcal{K}$ and we only need to determine if $\nabla^2 f(\mathbf{x})$ is PSD or not.*

In facts, many functions satisfy this condition. Some examples are $f(\mathbf{x}) = \mathbf{x}^T\mathbf{x}$, $f(\mathbf{x}) = 1$, etc.

Proposition 2.9.

1. *If f is convex and g is convex, then $f + g$ is convex.*
2. *If f is convex, then $\forall \alpha \geq 0$, αf is convex.*
3. *If f is convex and g is convex, then $\max\{f, g\}$ is convex.*
4. *If $g(\mathbf{x}, \mathbf{y})$ is jointly convex is $\mathbf{x}$, $\mathbf{y}$, then $f(\mathbf{x}) = \inf_{\mathbf{y}} g(\mathbf{x}, \mathbf{y})$ is convex.*

Definition 2.10 (Concave Function). *Let $\mathcal{K} \subset \mathbb{R}^n$ be a convex set and $f : \mathcal{K} \to \mathbb{R}$, then f is concave is $-f$ is convex.*

We can see that one example of concave function is *log* function.

Definition 2.11 (Lipschitz)**.** *Let $||\cdot||$ be a norm, $c \geq 0$ and $f : \mathbb{R}^n \to \mathbb{R}$. Then f is called $c-$Lipschitz with respect to $||\cdot||$ if $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,*

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq c||\mathbf{x} - \mathbf{y}||. \tag{2.7}$$

Intuitively, it means that the difference between function value is smaller than the difference between the norm of difference between two points times some constant c.

Claim 2.12. *If f is a differentiable function, then f is c-Lipschitz with respect to $||\cdot||$ if and only if $\forall \mathbf{x} \in \mathbb{R}^n$,*

$$||f(\mathbf{x})||_* \leq c \tag{2.8}$$

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

Proof. We first prove the 'if' part.

Let $||\cdot||$ be a norm on $\mathbb{R}^n$ and $c \geq 0$, let $||\nabla f(\mathbf{x})||_* \leq c \ \forall \mathbf{x} \in \mathbb{R}^n$. Then, by the mean value theorem, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, there exists $t \in [0,1]$ such that

$$f(\mathbf{x}) = f(\mathbf{y}) + \langle \nabla f((1-t)\mathbf{x} + t\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle.$$

Then, by Hölder's inequality,

$$\begin{aligned} |f(\mathbf{x}) - f(\mathbf{y})| &= |\langle \nabla f((1-t)\mathbf{x} + t\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle| \\ &\leq ||\nabla f((1-t)\mathbf{x} + t\mathbf{y})||_* ||\mathbf{x} - \mathbf{y}|| \\ &\leq c||\mathbf{x} - \mathbf{y}||. \end{aligned}$$

That is, f is *c-Lipschitz* with respect to $||\cdot||$.

We will next prove the 'only if' part.

Suppose that f is *c-Lipschitz* with respect to $||\cdot||$ and let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, then we have

$$\begin{aligned} \langle \nabla f(\mathbf{x}), \mathbf{y} \rangle &= \lim_{h \to 0} \frac{f(\mathbf{x} + h\mathbf{y}) - f(\mathbf{y})}{h} \\ &\leq \lim_{h \to 0} \frac{c||\mathbf{x} + h\mathbf{y} - \mathbf{x}||}{h} \\ &= c \lim_{h \to 0} \frac{h||\mathbf{y}||}{h} \\ &= c||\mathbf{y}||. \end{aligned}$$

Notice that $\langle \nabla f(\mathbf{x}), \mathbf{y} \rangle$ is the directional derivative of f at $\mathbf{x}$ in the direction of $\mathbf{y}$. Then,

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

Therefore, $\forall \mathbf{x} \in \mathbb{R}^n$, $||\nabla f(\mathbf{x})|| \leq c$. □

Theorem 2.13 (**Jensen's Inequality**). *content...*

Theorem 2.14 (**Young's Inequality**). *content...*

2.2 Bregman Divergence

Chapter 3

Deviation Bound

Chapter 4

Online Learning Algorithm

Chapter 5

Game Theory

Chapter 6

Boosting

Chapter 7

Online Convex Optimization