CS 7545

Machine Learning Theory

Professor Jacob Abernethy

Organized by Yiwen Chen

Contents

1 Math Review **3**

1.1 Positive Semidefinite and Positive Definite Matrices 3

1.2 Norm 4

2 Convex Analysis **6**

2.1 Convex Set and Convex Function 6

2.2 Bregman Divergence 9

3 Deviation Bound **12**

4 Online Learning Algorithm **17**

4.1 Online Learning and Halving Algorithm 17

4.2 Exponential Weights Algorithm 18

4.3 Perception 24

5 Game Theory **26**

6 Boosting **27**

7 Online Convex Optimization **28**

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). *Let $\mathcal{K} \in \mathbb{R}^n$ be a convex set, X be a random variable on $\mathcal{K}$ and $f : \mathcal{K} \to \mathbb{R}$ be a convex function, then*

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)] \tag{2.9}$$

where $\mathbb{E}$ denotes the expectation of a random variable.

Theorem 2.14 (Young's Inequality). *Let $p, q > 0$ and satisfy that $\frac{1}{p} + \frac{1}{q} = 1$. Then $\forall a, b > 0$,*

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}. \tag{2.10}$$

Proof. Since $-\log$ is a convex function, then $\forall a, b > 0$,

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

Then apply exp to both side and since exp function is monotonically increasing, we can have that

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

□

Definition 2.15 (Strongly Convex). *Let $\mathcal{K}$ be a convex set, $f : \mathcal{K} \to \mathbb{R}$ be a differentiable function and $\alpha > 0$. Then f os α-strongly convex with respect to $||\cdot||$ if $\forall \mathbf{x}, \mathbf{y} \in \mathcal{K}$,*

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{\alpha}{2}||\mathbf{x} - \mathbf{y}||^2. \tag{2.11}$$

Notice that if f is strongly convex, then it must be a convex function. In addition, a strongly convex function grows at least quadratically.

Claim 2.16. *Let f: dom(f) $\to \mathbb{R}$ be a twice differentiable function, then f is α-strongly convex if and only if $\forall \mathbf{x} \in$ dom(f),*

$$\nabla^2 f(\mathbf{x}) - \alpha I \succeq 0.$$

Definition 2.17 (Smooth). *Let f: dom(f) $\to \mathbb{R}$ be a differentiable function, let $\alpha > 0$, then f is α-smooth if $\forall \mathbf{x}, \mathbf{y} \in$ dom(f),*

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{\alpha}{2}||\mathbf{x} - \mathbf{y}||^2.$$

Notice that in contrast to strongly convex, a smooth function grows **at most** quadratically.

Claim 2.18. *We have a similar claim for smooth function to strongly convex function. That is, let f: dom(f) $\to \mathbb{R}$ be a twice differentiable function, then f is α-smooth if and only if $\forall \mathbf{x} \in$ dom(f), we have*

$$\nabla^2 f(\mathbf{x}) - \alpha I \preceq 0.$$

Notice that although smooth seems to be contrary to strongly convex, a function f can be both α-strongly convex and β-smooth for some α and β. We will give an example next.

Example 2.18.1. *Let $M \in \mathbb{R}^{n \times n}$ be a positive definite matrix and let $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T M \mathbf{x}$. Denote λ_{min} as the smallest eigenvalue of M and λ_{max} as the largest eigenvalue of M. Then we can prove easily that M is λ_{min}-strongly convex and λ_{max}-smooth.*

Proof. First, by taking derivative, it can be seen that $\nabla f(\mathbf{x}) = \mathbf{x}^T M$. Then, take the second derivative and we can get $\nabla^2 f(\mathbf{x}) = M$. By Claim 2.18 and Claim 2.16, we can show that f is λ_{min}-strongly convex and λ_{max}-smooth. □

2.2 Bregman Divergence

In this section, we will define Bregman Divergence and see its relation with convexity analysis.

Definition 2.19. *Given a function f: $\mathcal{U} \to \mathbb{R}$, we can define its Bregman Divergence as*

$$D_f(\mathbf{x}, \mathbf{y}) := f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle. \tag{2.12}$$

Example 2.19.1. *Let* $f(\mathbf{x}) = \frac{1}{2}||\mathbf{x}||_2^2$*, then*

$$D_f(\mathbf{x}, \mathbf{y}) = \frac{1}{2}||\mathbf{x} - \mathbf{y}||_2^2.$$

Notice that only in this situation is Bregman divergence a quadratic function.

Example 2.19.2. *Let* $\Delta^n = \{\mathbf{p} \in \mathbb{R}^n | \sum_{i=1}^n p_i = 1, p_i \geq 0\}$ *be the simplex with dimension n. Let* $f(\mathbf{p}) = \sum_{i=1}^n p_i \log p_i$ *be the entropy function and suppose* $f(0) = 0$*. Then*

$$D_f(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^n p_i \log \frac{p_i}{q_i} = KL(\mathbf{p}||\mathbf{q}),$$

where KL is Kullback-Leibler divergence.

We will next talk about some facts about Bregman divergence.

Fact 2.19.1.

1. *If* f *is convex, then* $D_f(\mathbf{x}, \mathbf{y}) = 0$ *if and only if* $\mathbf{x} = \mathbf{y}$*.*
2. f *is* μ*-strongly convex if and only if* $D_f(\mathbf{x}, \mathbf{y}) \geq \frac{\mu}{2}||\mathbf{x} - \mathbf{y}||^2$*.*
3. f *is* β*-smooth if and only if* $D_f(\mathbf{x}, \mathbf{y}) \leq \frac{\beta}{2}||\mathbf{x} - \mathbf{y}||^2$
4. *If* $D_f(\mathbf{x}, \mathbf{y}) = D_f(\mathbf{y}, \mathbf{x})$*, then* f *is quadratic.*
5. *Pinsker's Inequality:* $KL(\mathbf{p}, \mathbf{q}) \geq \frac{1}{2}||\mathbf{p} - \mathbf{q}||_1^2$*. It means that entropy function is* 1*-strongly convex in* $||\cdot||_1$*.*

Definition 2.20 (Fenchel Dual Conjugate). *Let* f *be convex, the Fenchel dual conjugate of* f *is defined as*

$$f^*(\theta) = \sup_{\mathbf{x} \in dom(f)} \left[\langle \mathbf{x}, \theta \rangle - f(\mathbf{x})\right]. \tag{2.13}$$

Claim 2.21. $f^*(\theta)$ *is convex.*

Proof. It can be seen that $g_{\mathbf{x}}(\theta) = \langle \mathbf{x}, \theta \rangle - f(\mathbf{x})$ is a linear function. Linear function is convex. In addition, $f^*(\theta) = \sup_{\mathbf{x} \in \mathrm{dom}(f)} g_{\mathbf{x}}(\theta)$ is the supreme of convex function, which means that f is also convex. □

Example 2.21.1. *If* $f(\mathbf{x}) = \frac{1}{2}||\mathbf{x}||_2^2$*, then* $f^*(\theta) = \frac{1}{2}||\theta||_2^2$*.*

Example 2.21.2. *If* $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T M \mathbf{x}$ *and* M *is a positive definite matrix,*

$$f^*(\theta) = \sup_{\mathbf{x} \in dom(f)} \left[\langle \mathbf{x}, \theta \rangle - \frac{1}{2}\mathbf{x}^T M \mathbf{x}\right].$$

Let $g(\mathbf{x}) = \langle \mathbf{x}, \theta \rangle - \frac{1}{2}\mathbf{x}^T M \mathbf{x}$*. Take the derivative over* $\mathbf{x}$ *and we can get that*

$$\begin{aligned} &\nabla_{\mathbf{x}} g(\mathbf{x}) = 0 \\ \Leftrightarrow\ &\theta - M\mathbf{x} = 0 \\ \Leftrightarrow\ &\mathbf{x} = M^{-1}\theta. \end{aligned}$$

Substitute the value of $\mathbf{x}$ *into* $f^*(\theta)$ *and we can get that*

$$\begin{aligned} f^*(\theta) &= \theta^T M^{-1}\theta - \frac{1}{2}(M^{-1}\theta)^T M M^{-1}\theta \\ &= \frac{1}{2}\theta^T M^{-1}\theta. \end{aligned}$$

Example 2.21.3. *Let* $f(x) = \frac{1}{p}||\mathbf{x}||_p^p = \frac{1}{p}\sum_{i=1}^{n} x_i^p$*, then*

$$f^*(\theta) = \frac{1}{q}||\theta||_q^q,$$

where $\frac{1}{p} + \frac{1}{q} = 1$.

There is one famous inequality for Fenchel dual conjugate called Fenchel-Yang Inequality.

Theorem 2.22 (Fenchel-Yang Inequality)**.** $\forall \mathbf{x} \in dom(f)$ *and* $\forall \theta \in dom(f^*)$*, it holds that*

$$f(\mathbf{x}) + f^*(\mathbf{x}) \geq \langle \mathbf{x}, \theta \rangle. \tag{2.14}$$

Proof. By the definition of Fenchel dual conjugate, $\forall \mathbf{x}$ and θ,

$$\begin{aligned} f^*(\theta) + f(\mathbf{x}) &= \sup_{\mathbf{y}} \left[\langle \mathbf{y}, \theta \rangle - f(\mathbf{y})\right] + f(\mathbf{x}) \\ &\geq \langle \mathbf{x}, \theta \rangle - f(\mathbf{x}) + f(\mathbf{x}) \\ &= \langle \mathbf{x}, \theta \rangle. \end{aligned}$$

□

Corollary 2.23. *By the inequality above, we can get directly that*

$$\frac{1}{p}||\mathbf{x}||_p^p + \frac{1}{q}||\theta||_q^q \geq \langle \mathbf{x}, \theta \rangle.$$

I will next discuss some facts about Fenchel dual.

Fact 2.23.1.

1. *If* f *is closed, then* $(f^*)^8 = f$.
2. $\nabla f(\nabla f^*(\theta)) = \theta$ *and* $\nabla f^*(\nabla f(\mathbf{x})) = \mathbf{x}$.
3. *If* f *is differentiable, then* $D_f(\mathbf{x}, \mathbf{y}) = D_{f^*}(\nabla f(\mathbf{y}), \nabla f(\mathbf{x}))$.
4. *If* f *is* μ*-strongly convex with respect to* $||\cdot||$*, then* f^* *is* $\frac{1}{\mu}$*-smooth with respect to* $||\cdot||_*$*, where* $||\cdot||_*$ *is dual to* $||\cdot||$.

Chapter 3

Deviation Bound

In this chapter several famous deviation bounds are mentioned. There bounds will be heavily used in later chapters.

First of all, some simple definitions from measure theory will be discussed.

Definition 3.1 (**Random Variable).** *A random variable X is a measurable function from a sigma algebra $\Omega \to \mathbb{R}$.*

The definition of measurable function and sigma algebra is included in measure theory, which will not be discussed here.

Definition 3.2 (**Cumulative Distribution Function).** *The CDF (cumulative distribution function) pf a random variable X is defined as*

$$F(t) := \Pr(X \leq t). \tag{3.1}$$

Definition 3.3 (**Probability Density Function).** *If $F(t)$ is differentiable, then the PDF (probability density function) is defined as*

$$f(t) = F'(t). \tag{3.2}$$

Definition 3.4 (**Variance).** *The variance of a random variable X is defined as*

$$Var(X) = \mathbb{E}[(X - \mu)]^2, \tag{3.3}$$

where $\mu = \mathbb{E}(X)$ is the expectation of X.

After the above definitions, we will next introduction some useful deviation bounds.

Theorem 3.5 (**Markov's Inequality).** *Let X be a random variable and $X \geq 0$, then $\forall t \geq 0$, we have that*

$$\Pr(X \geq t) \leq \frac{\mathbb{E}[X]}{t}. \tag{3.4}$$

Proof. The proof of above theorem is not very complex.

First of all, let $Z_t := \mathbf{1}(X \geq t) \cdot t$, where $\mathbf{1}(\cdot)$ is the indicator function.

Notice that $X \geq Z_t$ for all t. The reason is that if $X < t$, then $Z_t = 0 \leq X$. On the contrary, if $X \geq t$, then $Z_t = t \leq X$.

Use this, we can know that

$$\begin{aligned}\mathbb{E}(X) &\geq \mathbb{E}(Z_t) \\ &= \mathbb{E}[\mathbf{1}(X \geq t)] \cdot t \\ &= \Pr(X \geq t) \cdot t\end{aligned}$$

This directly shows that $\Pr(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$. □

Since $\Pr(X \geq t) = 1 - F(t)$, we can know from above theorem that $F(t) \geq 1 - \frac{\mathbb{E}(X)}{t}$.

Theorem 3.6 (Chebyshev's Inequality)**.** *Let $\mathbb{E}(X) = \mu$ and $Var(X) = \sigma^2$, then $\Pr(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$.*

It turns out that this theorem can be easily proved with Markov Inequality.

Proof.

$$\begin{aligned}\Pr(|X - \mu| \geq t) &= \Pr(|X - \mu|^2 \geq t^2) \\ &\leq \frac{\mathbb{E}[(X - \mu)]^2}{t^2} \\ &= \frac{\sigma^2}{t^2}.\end{aligned}$$

□

Next, we will define gaussian distribution (also called normal distribution).

Theorem 3.7 (Gaussian Distribution)**.** *Let's say $X \sim N(\mu, \sigma^2)$, it means that the pdf of X is*

$$P(X = x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x - \mu)^2}{2\sigma^2}).$$

This distribution is heavily used in many different areas.

Fact 3.7.1. *If $X \sim N(0, \sigma)$, and $t > 0$, then $\mathbb{E}[\exp(tX)] = \exp(\frac{t^2 s^2}{2})$.*

This can be proved just by taking integral of this pdf function.

Next, we will define subgaussian distribution, which will be used later in some inequalities.

Definition 3.8 (Subgaussian Distribution)**.** *A random variable X with $\mathbb{E}(X) = 0$ is a subgaussian distribution with respect to variance proxy σ^2 if*

$$\mathbb{E}[\exp(tX)] \leq \exp(\frac{t^2\sigma^2}{2}). \tag{3.5}$$

We will then give an example of subgaussian distribution.

Example 3.8.1. *Let X be a bounded random variable, that is, $a \leq X \leq b$ and $\mathbb{E}(X) = 0$. Then X is subgaussian with respect to variance proxy $\frac{(b-a)^2}{4}$, which means*

$$\mathbb{E}[\exp(tX)] \leq \exp(\frac{t^2(b - a)^2}{8}).$$

This is also called Hoeffiding's Lemma.

Claim 3.9. *If a random variable X is subgaussian with respect to variance proxy σ^2, then*

$$P(|X| > t) \leq 2\exp(-\frac{t^2}{2\sigma^2}). \tag{3.6}$$

Using subgaussian, we can introduce Hoeffding's inequality, which is extremely helpful in later material.

Theorem 3.10 (Hoeffding's Inequality). *Let $X_1, \cdots, X_n$ be independent random variables, with $\mathbb{E}(X_i) = \mu_i$ and $a_i \leq X_i \leq b_i$. Let $\bar{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\mu = \frac{1}{n}\sum_{i=1}^{n}\mu_i$. It holds that*

$$\Pr(|\bar{X_n} - \mu| > t) \leq 2\exp(-\frac{2n^2t^2}{\sum_{i=1}^{n}(a_i - b_i)^2}). \tag{3.7}$$

We will next give the proof of this theorem.

Proof. Notice that $\Pr(|\bar{X_n} - \mu| > t) \leq \Pr(\bar{X_n} - \mu > t) + \Pr(\bar{X_n} - \mu < -t)$. Then, we just need to prove that $\Pr(\bar{X_n} - \mu > t) \leq \exp(-\frac{2n^2t^2}{\sum_{i=1}^{n}(a_i - b_i)^2})$.

$$\begin{aligned}
\Pr(\bar{X_n} - \mu > t) &= \Pr(\exp(s(\bar{X_n} - \mu)) > \exp(st)) \\
\text{(Markov Inequality)} &\leq \frac{\mathbb{E}[\exp(s(\bar{X_n} - \mu))]}{\exp(st)} \\
&= \exp(-st)\mathbb{E}[\exp(s(\bar{X_n} - \mu))] \\
&= \exp(-st)\mathbb{E}[\Pi_{i=1}^{n}\exp(\frac{s}{n}(X_i - \mu_i))] \\
\text{(Independence)} &= \exp(-st)\Pi_{i=1}^{n}\mathbb{E}[\exp(\frac{s}{n}(X_i - \mu_i))] \\
(X_i - \mu_i \text{ Subgaussian}) &\leq \exp(-st)\Pi_{i=1}^{n}\exp(\frac{s^2}{8n^2}(b_i - a_i)^2) \\
&= \exp(\frac{s^2}{8n^2}\sum_{i=1}^{n}(a_i - b_i)^2 - st) \\
(s \text{ takes } \frac{4tn^2}{\sum_{i=1}^{n}(a_i - b_i)^2}) &= \exp(\frac{-2t^2n^2}{\sum_{i=1}^{n}(a_i - b_i)^2})
\end{aligned}$$

Similarly, we can prove that $\Pr(\bar{X_n} - \mu < -t) \leq \exp(\frac{-2t^2n^2}{\sum_{i=1}^{n}(a_i - b_i)^2})$. Combine the above two conclusions and we can get that $\Pr(|\bar{X_n} - \mu| > t) \leq 2\exp(\frac{-2t^2n^2}{\sum_{i=1}^{n}(a_i - b_i)^2})$. □

Hoeffding's Inequality is very useful in proving other theorems. To see that, let's use it to prove the following claim.

Claim 3.11. *Let $X_1, \cdots, X_n$ be independent random variable and $0 \leq X_i \leq 1$, then with probability at least $1 - \delta$, it holds that*

$$|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_i]| \leq \sqrt{\frac{\log 2/\delta}{2n}}. \tag{3.8}$$

Proof. Let $\delta = 2\exp(-\frac{2n^2t^2}{\sum_{i=1}^{n}(a_i-b_i)^2}) = 2\exp(-2nt^2)$ since $a_i = 0$ and $b_i = 1$. Then we can get that $t = \sqrt{\frac{\log 2/\delta}{2n}}$. By using Hoeffding's Inequality, we can get that

$$\Pr(|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_i]| \geq \frac{\log 2/\delta}{2n}) \leq \delta.$$

Therefore, we can get that $\Pr(|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_i]| \leq \frac{\log 2/\delta}{2n}) \geq 1 - \delta$. □

Afterwards, let's will introduction the concept of Martingale.

Definition 3.12 (Martingale). *A sequence of random variables $Z_0, Z_1, \cdots, Z_n$ is a Martingale sequence if the following two conditions are satisfied for any i*

1. $\mathbb{E}[|Z_i|] < \infty$
2. $\mathbb{E}[Z_i|Z_1, \cdots, Z_{i-1}] = Z_{i-1}$

Next let's see two examples of Martingale.

Example 3.12.1. *Let $X_1, \cdots, X_n$ be bounded i.i.d (independent and identically distributed) random variables, and $\mathbb{E}[X_i] = 0$. Let $Z_i = \sum_{j=1}^{i} X_j$, and set $Z_0 = 0$, then $Z_0, \cdots, Z_n$ is a martingale.*

Proof.

$$\begin{aligned}
\mathbb{E}[Z_i|Z_1, \cdots Z_{i-1}] &= \mathbb{E}[Z_i|X_1, \cdots, X_{i-1}] \\
&= \mathbb{E}[X_1 + \cdots + X_i|X_1, \cdots, X_{i-1}] \\
&= \mathbb{E}[X_1 + \cdots, X_{i-1}|X_1, \cdots, X_{i-1}] + \mathbb{E}[X_i|X_1, \cdots, X_{i-1}] \\
&= X_1 + \cdots X_{i-1} + \mathbb{E}[X_i] \\
&= Z_{i-1} + 0 \\
&= Z_{i-1}.
\end{aligned}$$

□

Example 3.12.2. *Let $X_1, \cdots, X_n$ be i.i.d, with $\mathbb{E}[X_i] = 0$ and $Var(X_i) = \sigma^2$. Let $S_i = \sum_{j=1}^{i} X_j$ and $Z_i = S_i^2 - i\sigma^2$. Then Z_i is a martingale.*

Proof. Notice that

$$\begin{aligned}
\mathbb{E}[X_n^2] &= Var(X_n) + \mathbb{E}[X_n]^2 \\
&= Var(X_n) \\
&= \sigma^2.
\end{aligned}$$

In addition,

$$\begin{aligned}
\mathbb{E}[S_{n-1}X_n|X_1, \cdots, X_{n-1}] &= \mathbb{E}[X_n]\mathbb{E}[S_{n-1}|X_1, \cdots, X_{n-1}] \\
&= 0.
\end{aligned}$$

Therefore, we have that

$$\begin{aligned}
\mathbb{E}[Z_n|X_1,\cdots,X_{n-1}] &= \mathbb{E}[S_n^2 - n\sigma^2|X_1,\cdots,X_{n-1}] \\
&= \mathbb{E}[(X_n + S_{n-1})^2 - n\sigma^2|X_1,\cdots,X_{n-1}] \\
&= \mathbb{E}[X_n^2 + S_{n-1}^2 + 2X_nS_{n-1} - n\sigma^2|X_1,\cdots,X_{n-1}] \\
&= \mathbb{E}[\sigma^2 + S_{n-1}^2 - n\sigma^2|X_1,\cdots,X_{n-1}] \\
&= S_{n-1}^2 - (n-1)\sigma^2 \\
&= Z_{n-1}.
\end{aligned}$$

□

Finally, we come to the last inequality of this chapter, which combines Hoeffding's Lemma and definition of Martingale.

Theorem 3.13 (Azuma's Inequality). *Let $Z_0, Z_1, \cdots, Z_n$ be a martingale, with $|Z_i - Z_{i-1}| \leq c_i$ for some $c_1, \cdots, c_n$. Then we have that*

$$\Pr(Z_n - Z_0 > t) \leq \exp(-\frac{t^2}{2\sum_{i=1}^{n} c_i^2}). \tag{3.9}$$

Proof.

$$\begin{aligned}
\Pr(Z_n - Z_0 > t) &= \Pr(\exp(s(Z_n - Z_0)) > \exp(st)) \\
&\leq \mathbb{E}[\exp(s(Z_n - Z_0))]\exp(-st) \\
&= \exp(-st)\mathbb{E}[\exp(s(Z_{n-1} - Z_0))\exp(s(Z_n - Z_{n-1}))] \\
&= \exp(-st)\mathbb{E}[\mathbb{E}[\exp(s(Z_{n-1} - Z_0))\exp(s(Z_n - Z_{n-1}))]|Z_0,\cdots,Z_{n-1}] \\
&= \exp(-st)\mathbb{E}[\exp(s(Z_{n-1} - Z_0))\mathbb{E}[\mathbb{E}[\exp(s(Z_n - Z_{n-1}))]|Z_0,\cdots,Z_{n-1}]
\end{aligned}$$

Since $|Z_n - Z_{n-1}| \leq c_n$, we can get that $Z_n - Z_{n-1} \in [-c_n, c_n]$. In addition, we know that $\mathbb{E}[Z_n - Z_{n-1}|Z_0,\cdots,Z_{n-1}] = Z_{n-1} - Z_{n-1} = 0$. Therefore, by using Hoeffding's Lemma, we know that

$$\begin{aligned}
\Pr(Z_n - Z_0 > t) &\leq \exp(-st)\mathbb{E}[\exp(s(Z_{n-1} - Z_0))\exp(\frac{s^2c_n^2}{2}) \\
&= \exp(-st + \frac{s^2c_n^2}{2})\mathbb{E}[\exp(s(Z_{n-1} - Z_0)) \\
&\leq \exp(-st + \frac{s^2c_n^2}{2})\exp(\frac{s^2c_{n-1}^2}{2})\mathbb{E}[\exp(s(Z_{n-2} - Z_0)) \\
&\cdots \\
&\leq \exp(-st + \frac{s^2}{2}\sum_{i=1}^{n} c_i^2)
\end{aligned}$$

Let $s = \frac{t^2}{\sum_{i=1}^{n} c_i^2}$ and we can get that $\Pr(Z_n - Z_0 > t) \leq \exp(-\frac{t^2}{2\sum_{i=1}^{n} c_i^2})$. □

We can see that the above proof is a little bit similar to the proof of Hoeffding's Inequality in that they both take the exponential of both sides and then use Markov Inequality to transform probability into expectation. Afterwards, they perform some kind of relaxation for the expectation.

Chapter 4

Online Learning Algorithm

In this chapter we will talk about the basics of online learning algorithm step by step. We will first introduce the online learning setting and an algorithm called halving algorithm, which is simple but has strong assumption. Afterwards, we will relax the assumption and introduction other algorithms such like exponential weighted algorithm.

4.1 Online Learning and Halving Algorithm

First of all, let's introduce a simple setting of online learning. In this setting,

- Goal: predict rain/shine.
- Have a set of N weather experts.
- On each day t, expert i predicts $x_i^t \in \{0, 1\}$.
- Based on the predictions of experts, the algorithm predicts $\hat{y^t} \in \{0, 1\}$.
- Nature reveals $y^t \in \{0, 1\}$.
- The number of mistakes increases by one if $\hat{y^t} \neq y^t$.
- Assume there exists an perfect expert j, such that $x_j^t = y^t$ for all t.

We hope to develop an algorithm $\mathcal{A}$ such that we can get an upper bound of total number of mistakes. In this setting, we can find a simple yet good algorithm called Halving algorithm such that the total number of mistakes is no larger than $\log N$.

Algorithm 1 Halving Algorithm

Let a set $C_1 = \{1, 2, \cdots, N\}$.
For $t = 1, 2, \cdots$
    Observe x_i^t, $\forall i \in C_t$.
    $\hat{y^t} = round(\frac{1}{|C_t|} \sum_{i \in C_t} x_i^t)$
    Let $C_{t+1} = C_t$
    For all $i \in C_t$
        If $x_i^t \neq y_t$, remove i from C_{t+1}.

Basically, the algorithm will output the round of the average over all experts' output in the set. Afterwards, it will remove all the experts whose output is not the same as ground truth and continue. For this algorithm, we can have the following claim.

Claim 4.1. *For halving algorithm, it satisfies that the total number of mistakes is no larger than* $\log_2 N$.

Proof. The proof is quite simple. Initially we have $|C_1| = N$. Each time the number of mistakes increases, we can have that $|C_{t+1}|/|C_t| \leq 1/2$ since more than half of experts produces the different prediction than ground truth.

Therefore, if the number of mistakes reaches $\log_2 N$, we will have that $|C_t| = 1$, which means the only expert in the set is the perfect expert and no more mistake will be made. □

Next, let's see an example of using halving algorithm to solve some problem.

Example 4.1.1 (Betting on sports)**.**

Problem setting: let's say there are n sport teams. On each round t, two teams i_t, j_t *play a match. Algorithm* $\mathcal{A}$ *needs to predict if* i_t *beats* j_t *or vice versa. Then games happens, either* i_t *or* j_t *wins.*

Assume there exists a permutation $\pi^* \in S_n$, i_t *beats* j_t *if and only if* $\pi^*(i_t) > \pi^*(j_t)$.

Now we need an algorithm that can minimize the number of mistakes to gain more revenue. We can design our algorithm by reducing to halving algorithm as following:

Algorithm 2 Betting Algorithm

Treat every $\pi \in S_n$ as an expert.
Let the prediction of expert i at round t be $x_i^t = \mathbf{1}[\pi(i_t) > \pi(j_t)]]$.
Let the output of nature as $y_t = \mathbf{1}[\pi^*(i_t) > \pi^*(j_t)]$.
Run halving algorithm to the set of experts.

By halving algorithm, we know that the total number of mistakes is no larger than $\log_2 |S_n|$. *In this case it is* $\log_2 n! = O(n \log n)$.

Halving algorithm is very intuitive and the upper bound is also very satisfying. However, the problem is that it assumes that there is a perfect experts that will not make any mistakes. This assumption is too strong and in most cases it is not satisfied. To handle this issue, in next section we introduction exponential weights algorithm. It removes the perfect expert assumption thus can generalize to more situations.

4.2 Exponential Weights Algorithm

In this section, we change the assumption. We no longer assume that there exists a perfect expert, and we introduce two new notations.

- $M_T(i) = \sum_{i=1}^{T} \mathbf{1}[x_i^t \neq y^t]$
- $M_T(\mathcal{A}) = \sum_{i=1}^{T} \mathbf{1}[\hat{y^t} \neq y^t]$

In the other words, $M_T(i)$ is the number of mistakes expert i makes up to time T while $M_T(\mathcal{A})$ is the number of mistakes the algorithm makes.

Since we have remove the assumption that there is an expert who will never make mistake, we need a smoother algorithm than halving algorithm because in halving algorithm we are zero-tolerant to any mistake. Before talking about exponential weighted algorithm, let's first talk about weighted majority algorithm.

Algorithm 3 Weighted Majority Algorithm

Let $w_i^1 = 1$, for $i = 1, \cdots, N$.
Let $\epsilon \in (0, 1)$ be a parameter we choose.
For $t = 1, 2, \cdots$

Algorithm predicts $\hat{y^t} = round(\frac{\sum_{i=1}^{N} w_i^t x_i^t}{\sum_{i=1}^{N} w_i^t})$

For $i = 1, \cdots, N$

$w_i^{t+1} = w_i^t (1-\epsilon)^{\mathbf{1}[x_i^t \neq y^t]}$

Intuitively, different from halving algorithm, in weighted majority algorithm, if an expert makes a mistake, we will reduce its weight rather than remove it from the set. The expert will still have contribution to the algorithm but the weight is smaller. One can also see that halving algorithm is a special case of weighted majority algorithm when $\epsilon = 1$.

For this algorithm, we have the following theorem:

Theorem 4.2. *For any ϵ and expert i, no matter what the sequence $y^1, \cdots, y^T$ is, we can have that*

$$M_T(WMA) \leq \frac{2 \log_e N}{\epsilon} + 2(1+\epsilon) M_T(i). \tag{4.1}$$

To prove this theorem, we will need the following lemma first:

Lemma 4.3. *The following inequalities are valid.*

1. $\log(1+x) \leq x$.
2. $1 + x \leq \exp(x)$.
3. $\exp(\alpha x) \leq 1 + (\exp(\alpha) - 1)x$, *for* $x \in [0, 1]$.
4. $-\log(1+x) \leq -x + x^2$, *for* $x \in [-1, \frac{1}{2}]$.

There are many kinds of proof to this lemma online, thus we will omit the proof here. Next, we will go directly into the proof to this theorem.

Proof. Let $\Phi_t = \sum_{i=1}^{N} w_i^t$. Notice that $\Phi_1 = N$. Furthermore, since all weights are non-negative, we can have that $\Phi_{T+1} \geq w_i^{T+1} = (1-\epsilon)^{M_T(i)}$.

We now take a look at the case when WMA makes a mistake at round t. In that case, we know that at least half of total weights are wrong. That is, at least $\frac{\sum_{i=1}^{N} w_i^t}{2}$ will shrink. By the updating

formula of weights in WMA, we can have that

$$\begin{aligned}\Phi_{t+1} &\leq \frac{\sum_{i=1}^{N} w_i^t}{2} + (1-\epsilon)\frac{\sum_{i=1}^{N} w_i^t}{2} \\ &= (1-\frac{\epsilon}{2})\sum_{i=1}^{N} w_i^t \\ &= (1-\frac{\epsilon}{2})\Phi_t.\end{aligned}$$

Therefore, $\Phi_{T+1} \leq \Phi_0(1-\frac{\epsilon}{2})^{M_T(WMA)} = N(1-\frac{\epsilon}{2})^{M_T(WMA)}$.
Now, combine above inequalities and we can have that

$$(1-\epsilon)^{M_T(i)} \leq \Phi_{T+1} \leq N(1-\frac{\epsilon}{2})^{M_T(WMA)}.$$

Take log on both left side and right side we can get that

$$M_T(i)\log(1-\epsilon) \leq \log(N) + M_T(WMA)\log(1-\frac{\epsilon}{2}).$$

Use inequality 1 and 4 in the lemma and we get

$$M_T(i)(\epsilon+\epsilon^2) \geq -\log(N) + \frac{\epsilon}{2}M_T(WMA).$$

This directly indicates that

$$M_T(WMA) \leq \frac{2\log N}{\epsilon} + 2(1+\epsilon)M_T(i).$$

□

Since i can take any value, we know that $M_T(WMA) \leq \frac{2\log N}{\epsilon} + 2(1+\epsilon)M_T(i^*)$ where i^* is the best expert that makes least number of mistakes

Following this theorem, by setting $\epsilon = \sqrt{\frac{\log N}{M_T(i^*)}}$, we reach that

$$\begin{aligned}M_T(WMA) &\leq \frac{2\log N}{\sqrt{\frac{\log N}{M_T(i^*)}}} + 2(1+\sqrt{\frac{\log N}{M_T(i^*)}})M_T(i^*) \\ &= 2\sqrt{M_T(i^*)\log N} + 2M_T(i^*) + 2\sqrt{\frac{\log N}{M_T(i^*)}}M_T(i^*) \\ &= 2M_T(i^*) + 4\sqrt{M_T(i^*)\log N}.\end{aligned}$$

This concludes our discussion about weighted majority algorithm. We will now start discussing about exponential weighted algorithm. Before talking about the algorithm details, let's first introduce two new settings for this algorithm.

- Setting 1: Continuous Prediction
  - At time t, each expert i predicts $x_i^t \in [0,1]$.

- Algorithm predicts $\hat{y^t} \in [0,1]$.
- Nature outcomes $y^t \in \{0,1\}$.
- We have a convex loss function $l : [0,1] \times \{0,1\} \to \mathbb{R}$.
- The loss of expert i at t is $l(x_i^t, y^t)$.
- The loss of algorithm at t is $l(\hat{y^t}, y^t)$.
- Define $L_t(i) = \sum_{s=1}^{t} l(x_i^s, y^s)$.
- Define $L_t(Alg) = \sum_{s=1}^{t} l(\hat{y}^s, y^s)$.

We can see that $L_t(i) = L_{t-1}(i) + l(x_i^t, y^t)$. We can give some examples of loss functions.

1. Absolute loss: $l(\hat{y}, y) = |\hat{y} - y|$.
2. Square loss: $l(\hat{y}, y) = (\hat{y} - y)^2$.
3. Log loss: $l(\hat{y}, y) = -y \log \hat{y} - (1-y) \log(1-\hat{y})$.

- Setting 2: Hedge/Action Setting
  - There are N actions.
  - Algorithm must (randomly) select an action i_t on day t.
  - This is equivalent that algorithm selects a distribution $\mathbf{p}^t \in \Delta_N$.
  - Nature chooses losses $l^t = [l_1^t, \cdots, l_N^t] \in [0,1]^N$, where l_i^t means the cost of choosing i.
  - Then, the expected loss to this algorithm is defined as $p^t l^t = \mathbb{E}[l_i^t]$.
  - Define $L_t(i) = \sum_{s=1}^{t} l_i^s$.
  - Define $L_t(Alg) = \sum_{s=1}^{t} p^s l^s$.

To continue, we define the regret of an algorithm, which is the cost function that we want to minimize.

Definition 4.4 (Regret). *The regret of an algorithm is defined as*

$$Regret_T(Alg) = L_T(Alg) - \min_{i \in [N]} L_T(i). \tag{4.2}$$

Now we will give the algorithm details of Exponential Weights Algorithm:

Algorithm 4 Exponential Weights Algorithm

Let $w_i^1 = 1$, for $i = 1, \cdots, N$. Choose $\eta > 0$.
For $t = 1, \cdots, T$
  For $i = 1, \cdots, N$
    $w_i^t = \exp(-\eta L_t(i))$
  Prediction Setting:
    Observe $x_1^t, \cdots, x_N^t$
    Output $\hat{y^t} = (\sum_{i=1}^{N} w_i^t x_i^t)/(\sum_{i=1}^{N} w_i^t)$
    Observed y^t.
  Hedge Setting:
    Output $\mathbf{p}^t = \frac{1}{\sum_{i=1}^{N} w_i^t}[w_1^t, \cdots, w_N^t]$
    Observed $l^t \in [0,1]^N$.

For prediction setting, we can see that it is the same as WMA algorithm except the way it updates the weights. We have the following theorem for Exponential Weights Algorithm:

Theorem 4.5. *For any sequence of inputs and any choice of η, we have that*

$$L_T(EWA) \leq \frac{\eta L_t(i) + \log N}{1 - \exp(-\eta)} \tag{4.3}$$

Corollary 4.6. *For an excellent choice of η, we have that*

$$L_T(EWA) - L_T(i) \leq \log N + 2\sqrt{L_T(i^*) \log N}. \tag{4.4}$$

It means that the regret of algorithm is not larger than $\log N + 2\sqrt{L_T(i^*) \log N}$.

Compare it with the result of weighted majority algorithm, we can see that EWA has much better performance when $M_T(i^*) \gg \log N$. Next, before proving the theorem, let's first prove a lemma that will be used.

Lemma 4.7. *Let X be a random variable in $[0, 1]$, then $\log(\mathbb{E}[\exp(sX)]) \leq (\exp(s) - 1)\mathbb{E}[X]$.*

Proof. With the inequalities in Lemma 4.3, we can show that

$$\begin{aligned} \log(\mathbb{E}[\exp(sX)]) &\leq \log(\mathbb{E}[1 + (\exp(s) - 1)X]) \\ &=\leq \log(1 + (\exp(s) - 1)\mathbb{E}[X]) \\ &\leq ((\exp(s) - 1)\mathbb{E}[X]) \end{aligned}$$

□

Next, let's prove theorem 4.5. We will prove the algorithm in Hedge setting, but the proof is almost the same for prediction setting.

Proof. First of all, let's define a random variable X_t as $X_t = l(x_i^t, y^t)$ with probability $\frac{w_i^t}{\sum_{i=1}^{N} w_j^t}$.

Then, like in the proof of WMA, we also define a function $\Phi_t = -\log(\sum_{i=1}^{N} w_i^t)$. We can get that

$$
\begin{aligned}
\Phi_{t+1} - \Phi_t &= -\log(\frac{\sum_{i=1}^{N} w_i^{t+1}}{\sum_{i=1}^{N} w_i^t}) \\
&= -\log(\frac{\sum_{i=1}^{N} w_i^t \exp(-\eta l(x_i^t, y^t))}{\sum_{i=1}^{N} w_i^t}) \\
&= -\log(\sum_{i=1}^{N} \frac{w_i^t}{\sum_{j=1}^{N} w_j^t} \exp(-\eta l(x_i^t, y^t))) \\
&= -\log(\mathbb{E}[\exp(-\eta X_t)]) \\
&\geq (1 - \exp(-\eta))\mathbb{E}[X_t] \\
&= (1 - \exp(-\eta))(\sum_{i=1}^{N} \frac{w_i^t}{\sum_{j=1}^{N} w_j^t} l(x_i^t, y^t))) \\
(Jensen's\ Inequality) &\geq (1 - \exp(-\eta)) l(\sum_{i=1}^{N} \frac{w_i^t x_i^t}{\sum_{j=1}^{N} w_j^t}, y^t) \\
&= (1 - \exp(-\eta)) l(\hat{y^t}, y^t).
\end{aligned}
$$

In addition, we know that $\Phi_1 = -\log N$ and $\Phi_{T+1} = -\log(\sum_{i=1}^{N} \exp(-\eta L_T(i))) \leq \eta L_T(i)$ for any i. Therefore,

$$
\begin{aligned}
\log N + \eta L_T(i) &\geq \Phi_T - \Phi_1 \\
&= \sum_{t=1}^{T} (\Phi_{t+1} - \Phi_t) \\
&= \sum_{t=1}^{T} (1 - \exp(-\eta)) l(\hat{y^t}, y^t) \\
&= (1 - \exp(-\eta)) L_T(EWA).
\end{aligned}
$$

Therefore, we can have that for any i,

$$
L_T(EWA) \leq \frac{\log N + \eta L_T(i)}{1 - \exp(-\eta)}.
$$

□

From corollary 4.6, we know that by choosing η carefully, we can have that $Regret_T(EWA) \leq \log N + 2\sqrt{L_T(i^*) \log N}$. If the loss function $l(x_i^s, y^s) \in [0, 1]$, it holds that

$$
\begin{aligned}
\frac{L_T(EWA) - L_T(i)}{T} &\leq \frac{1}{T}(\log N + 2\sqrt{L_T(i^*) \log N}) \\
&= O(\frac{1}{\sqrt{T}}).
\end{aligned}
$$

It shows that $\frac{Regret_T}{T}$ will goes to zero as $T \to \infty$.

Hedge setting and prediction setting belong to a category called full information setting. It means besides the choice algorithm makes, it also knows the loss of all alternative choices. Next we will talk about Perception Algorithms.

4.3 Perception

We now introduce a new setting called linear prediction setting:

- At time t, observe $x^t \in \mathbb{R}^d$ with $||x^t||_1 \leq 1$.
- Define linear predictor as a function $h_w(\cdot)$ parameterized by $w \in \mathbb{R}^d$ and $h_w(\cdot) = sign(w \cdot x)$.
- Predict $\hat{y^t} \in \{-1, 1\}$ using some linear predictor.
- Outcome $y^t \in \{-1, 1\}$.

We assume that for some γ, $\exists w$ such that $||w||_2 \leq 1$ and $(w \cdot x^t)y^t > \gamma$ for any t. This is equivalent to $\exists w$, such that $||w||_2^2 \leq \frac{1}{\gamma^2}$ and $(w \cdot x^t)y^t > 1$ for any t.

Intuitively, this assumption means that there exists a perfect linear predictor with respect to margin γ. Based on this assumption, we can start talking about perception algorithm.

Algorithm 5 Perception

Let $w^1 = 0 \in \mathbb{R}^d$
For $t = 1, \cdots, T$
  $\hat{y^t} = sign(w^t \cdot x^t)$.
  Observe y^t.
  If $y^t(w^t \cdot x^t) > 0$, then $w^{t+1} = w^t$.
  Else, $w^{t+1} = w^t + x^t y^t$.

This algorithm is the same as gradient descent algorithm with loss function $l(w; (x, y)) = \max\{0, -(w \cdot x)y\}$, and

$$w^{t+1} = w^t - \nabla l(w^t; (x^t, y^t)).$$

For Perception Algorithm, the following theorem holds that

Theorem 4.8. *Let* $M_T = \sum_{i=1}^{T} \mathbf{1}[y^t(w^t \cdot x^t) < 0]$. *Assume* $\exists w^*$ *such that* $||w^*|| \leq \frac{1}{\gamma}$ *and* $(w^* \cdot x^t)y^t \geq 1$, *for any* t. *Then*

$$M_T \leq \frac{1}{\gamma^2}. \tag{4.5}$$

Proof. Suppose w^* satisfies the assumption in theorem, then define $\Phi_t = ||w^* - w^t||^2$.

Notice that

$$\begin{aligned} \Phi_1 = ||w^* - w^1||^2 &= ||w^*|| \\ &\leq \frac{1}{\gamma^2}. \end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{1}{\gamma^2} &\geq \Phi_1 - \Phi_{T+1} \\
&= \sum_{t=1}^{T} (\Phi_t - \Phi_{t+1}) \\
&= \sum_{t=1}^{T} (||w^* - w^t||^2 - ||w^* - w^{t+1}||^2) \\
&= \sum_{t:\text{mistake}(t)} (||w^* - w^t||^2 - ||w^* - w^t - x^t y^t||^2) \\
&= \sum_{t:\text{mistake}(t)} 2(w^* - w^t)(x^t y^t) - (y^t)^2 ||x^t||^2 \\
&= \sum_{t:\text{mistake}(t)} 2(w^* - w^t)(x^t y^t) - ||x^t||^2 \\
&\geq \sum_{t:\text{mistake}(t)} 2(w^* - w^t)(x^t y^t) - 1 \\
(y^t(w^t \cdot x^t) < 0) &\geq \sum_{t:\text{mistake}(t)} 2(w^*)(x^t y^t) - 1 \\
(y^t(w^* \cdot x^t) \geq 1) &\geq \sum_{t:\text{mistake}(t)} 1 \\
&= M_T.
\end{aligned}$$

Therefore, we can get directly that $M_T \leq \frac{1}{\gamma^2}$. □

Chapter 5

Game Theory

Chapter 6

Boosting

Chapter 7

Online Convex Optimization