

# Machine Learning Theory Notes

Gaotang Li

February 7, 2023

# Contents

<b>I</b>	<b>Generalization Theory</b>	<b>2</b>
<b>1</b>	<b>Supervised Learning Framework</b>	<b>3</b>
1.1	Basic Setups . . . . .	3
1.2	Empirical Risk Minimization . . . . .	3
1.3	Questions . . . . .	4
<b>2</b>	<b>Concentration Inequality</b>	<b>5</b>
2.1	Chebyshev's Inequality . . . . .	5
2.2	Hoeffding Inequality . . . . .	6
2.3	Bounded Difference Concentration Inequality . . . . .	7
<b>3</b>	<b>Rademacher Complexity</b>	<b>8</b>
3.1	Uniform Convergence . . . . .	8
3.2	Rademacher Complexity . . . . .	9

## Part I

# Generalization Theory

# Chapter 1

## Supervised Learning Framework

### 1.1 Basic Setups

In a supervised learning problem, we have a goal to predict a label given an input. Let  $S$  denote the dataset  $\{(x_i, y_i)\}_{i=1}^n$  for

- $x_i \in \mathcal{X}$ , the inputs in the input space.
- $y_i \in \mathcal{Y}$ , the label associated with  $x_i$  in the label space.

We assume that the data are drawn **i.i.d.** from an unknown probability distribution  $\mathcal{P}$  over  $\mathcal{X} \times \mathcal{Y}$ . We aim at learning a function mapping  $h: \mathcal{X} \rightarrow \mathcal{Y}$  (aka *hypothesis*, *predictor*, *model*).

To evaluate the performance of  $h$ , we specify a loss function. A loss function  $\ell: \mathcal{Y}, \mathcal{Y} \rightarrow \mathbb{R}$  measures the difference between the predicted label and the groundtruth label.

**Definition 1.1.1 (population risk).** The *population risk* of a *hypothesis*  $h$  is its expected loss over the data distribution  $\mathcal{P}$ :  $L_{\mathcal{P}}(h) = \mathbb{E}_{(x,y) \sim \mathcal{P}}[\ell(h(x), y)]$

**Example.** Examples of Loss Functions

- Classification: 0-1 loss  $\ell(\hat{y}, y) = \mathbb{1}(\hat{y} \neq y)$ .
- Regression: squared loss  $\ell(\hat{y}, y) = (\hat{y} - y)^2$

It is often impossible to consider all possible function mappings from  $\mathcal{X} \rightarrow \mathcal{Y}$ . We usually only consider a *hypothesis class*  $\mathcal{H}$ .

**Example.** Examples of  $\mathcal{H}$ .

- Linear function class:  $\mathcal{H} = \{h_{\theta} | h_{\theta}(x) = \theta^T x, \theta \in \mathbb{R}\}$
- General parametric function class:  $\mathcal{H} = \{h_{\theta} | h_{\theta}(x) = f(x, \theta), \theta \in \mathbb{R}^p\}$

### 1.2 Empirical Risk Minimization

**Definition 1.2.1 (Empirical Risk).** The *empirical risk* of a hypothesis  $h$  is its average loss over the dataset  $S$

$$L_s(h) = \frac{1}{n} \sum_{i=1}^n \ell(h(x_i), y_i)$$

Empirical risk minimization (ERM) is any algorithm that minimizes the empirical risk over the hypothesis class  $\mathcal{H}$ . We denote a hypothesis returned by ERM as  $\hat{h}_{ERM}$ , i.e.:

---

$$\hat{h}_{ERM} \in \operatorname{argmin}_{h \in \mathcal{H}} L_S(h)$$

If we assume  $h$  is independent of  $S$ , then  $\mathbb{E}_S[L_S(h)] = L_P(h)$ . But in reality  $h$  and  $S$  are not independent.

### 1.3 Questions

In the supervised learning part of this course, we are mainly interested in the following two fundamental problems:

- **Statistical:** What guarantee do we have about  $L_P(\hat{h}_{ERM})$ ?
- **Optimization:** When may ERM be achieved efficiently?

## Chapter 2

# Concentration Inequality

Concentration inequalities are a mathematical tool to study the relation between population and empirical quantities. Consider the following main question: for i.i.d. random variables  $X_1, \dots, X_n$ , how does  $\frac{1}{n} \sum_{i=1}^n X_i$  relate to  $\mathbb{E}[\frac{1}{n} \sum_{i=1}^n X_i] = \mu$ ?

### 2.1 Chebyshev's Inequality

**Lemma 2.1.1** (Markov's Inequality). Let  $X$  be a non-negative random variable, then for all  $t > 0$ ,

$$\Pr(X \geq t) \leq \frac{\mathbb{E}(x)}{t}$$

**Proof.**

$$\mathbb{E}(X) \geq \Pr(X < t) * 0 + \Pr(X > t) * t$$

■

**Theorem 2.1.1** (Chebyshev's Inequality). Let  $X$  be a random variable with finite expected value  $\mu$  and finite non-zero variance  $\sigma^2$ . Then for any real number  $t > 0$ ,

$$\Pr[|X - \mathbb{E}(X)| \geq t] \leq \frac{\text{Var}(X)}{t^2}$$

**Proof.**

$$\begin{aligned} \Pr[X - \mathbb{E}[X] \geq t] &= \Pr[(X - \mathbb{E}[X])^2 \geq t^2] \\ &= \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{t^2} \\ &= \frac{\text{Var}[X]}{t^2} \end{aligned}$$

■

**Remark.**

$$\Pr[|X - \mathbb{E}[X]| \geq t] \leq \frac{\mathbb{E}[|X - \mathbb{E}[X]|^P]}{t^P}$$

**Corollary 2.1.1.** Let  $x_1, \dots, x_n$  be i.i.d. random variables such that  $\mathbb{E}[x_i] = \mu$ ,  $\text{Var}[x_i] = \sigma^2$ . Then:

$$\Pr\left[\left|\frac{1}{n} \sum_{i=1}^n x_i - \mu\right| \geq t\right] \leq \frac{\sigma^2}{nt^2}$$

## 2.2 Hoeffding Inequality

**Lemma 2.2.1.** If  $X \in [0, 1]$  a.s. Then,

$$\mathbb{E} \left[ e^{\lambda(X - \mathbb{E}[X])} \right] \leq e^{\frac{\lambda^2}{8}}$$

for all  $\lambda \in \mathbb{R}$ .

**Proof.** Let  $Z = X - \mathbb{E}[X]$ , then  $\mathbb{E}[Z] = 0$ .

Define  $\psi(\lambda) := \log \mathbb{E} [e^{\lambda Z}]$ .

Using the Taylor expansion to get that  $\psi(\lambda) = \psi(0) + \lambda\psi'(0) + \frac{\lambda^2}{2}\psi''(\lambda')$  where  $\lambda'$  is between 0 and  $\lambda$ .

Here the first term  $\psi(0) = \log 1 = 0$ , and the second term  $\lambda\psi'(0) = \mathbb{E}[Z] = 0$ . The only thing we need to is to compute the third term. The idea is to bound the third term by  $1/4$ .

Then

$$\begin{aligned} \psi'(\lambda) &= \frac{\mathbb{E} [e^{\lambda Z} Z]}{e^{\lambda Z}} = \mathbb{E}[Y] \\ \psi''(\lambda) &= \frac{\mathbb{E} [e^{\lambda Z} Z^2]}{e^{\lambda Z}} - \left( \frac{\mathbb{E} [e^{\lambda Z} Z]}{e^{\lambda Z}} \right)^2 \\ &= \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2 \\ &= \text{Var}[Y] \end{aligned}$$

Where we can think of  $Y$  as a reweighted version of  $Z$ , and we have that

$$dP_Y(x) = \frac{e^{\lambda x}}{\mathbb{E} [e^{\lambda Z}]} dP_Z(x)$$

not finished yet...

■

**Remark.** We also call such random variables **subgaussian** random variables. Another interpretation is that bounded random variables are subgaussian.

Another reminder is that the expectation is in the form of **Moment Generating Function**, where  $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$ .

**Theorem 2.2.1 (Hoeffding Inequality).** Let  $X_1, \dots, X_n$  be i.i.d. random variables such that for each  $i$ ,  $X_i \in [0, 1]$  a.s. Then for all  $t > 0$ :

$$\begin{aligned} \Pr \left[ \frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n X_i \right] \geq t \right] &\leq e^{-2nt^2} \\ \Pr \left[ \frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n X_i \right] \leq -t \right] &\leq e^{-2nt^2} \end{aligned}$$

**Proof.** Let us use  $\bar{X}$  to denote  $\frac{1}{n} \sum_{i=1}^n X_i$ . Then we have that

$$\begin{aligned}
\Pr [\bar{X} - \mathbb{E}[\bar{X}] \geq t] &= \Pr \left[ e^{\lambda(\bar{X} - \mathbb{E}[\bar{X}])} \geq e^{\lambda t} \right] && \lambda > 0 \\
&\leq \frac{\mathbb{E} \left[ e^{\lambda(\bar{X} - \mathbb{E}[\bar{X}])} \right]}{e^{\lambda t}} && \text{Markov} \\
&= e^{-\lambda t} \cdot \mathbb{E} \left[ e^{\frac{\lambda}{n} \sum_{i=1}^n (X_i - \mathbb{E}[X_i])} \right] \\
&= e^{-\lambda t} \cdot \mathbb{E} \left[ \prod_{i=1}^n e^{\frac{\lambda}{n} (X_i - \mathbb{E}[X_i])} \right] \\
&= e^{-\lambda t} \prod_{i=1}^n \mathbb{E} \left[ e^{\frac{\lambda}{n} (X_i - \mathbb{E}[X_i])} \right] && \text{Independence} \\
&\leq e^{-\lambda t} \left( e^{\frac{1}{8} \left( \frac{\lambda}{n} \right)^n} \right) \\
&= e^{-\lambda t + \frac{\lambda^2}{8n}} \\
&= e^{-2nt^2} && \text{let } \lambda = 4nt
\end{aligned}$$

By symmetry, we complete the proof.  $\blacksquare$

**Remark (Equivalent Definition of Hoeffding Inequality).** Let  $X_1, \dots, X_n \in [0, 1]$  a.s. and independent,

$$\begin{aligned}
\forall \delta \in (0, 1), \text{ w.p. } \geq 1 - \delta: \quad & \frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n X_i \right] \leq \sqrt{\frac{\log \frac{1}{\delta}}{2n}} \\
\text{w.p. } \geq 1 - \delta: \quad & \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n X_i \right] - \frac{1}{n} \sum_{i=1}^n X_i \leq \sqrt{\frac{\log \frac{1}{\delta}}{2n}}
\end{aligned}$$

## 2.3 Bounded Difference Concentration Inequality

We are concerned with diffeomorphism (*i.e.* change in one coordinate) and formally

$$f(X_1, \dots, X_n) \mapsto \mathbb{E}[f(X_1, \dots, X_n)]$$

**Theorem 2.3.1 (Mcdiarmid's inequality).** Suppose  $X_1, \dots, X_n$  are independent random variables taking values in a set  $A$ . Let  $f: A^n \rightarrow \mathbb{R}$  be a function that satisfies the *bounded difference* condition:

$$\exists c_1, \dots, c_n > 0 \text{ s.t. } \forall x_1, \dots, x_n \in A, x'_i \in A |f(x_1, \dots, x_n) - f(x_1, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_n)| \leq c_i$$

Then, for all  $t > 0$ ,

$$\Pr [f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)] \geq t] \leq e^{-\frac{2t^2}{\sum_{i=1}^n c_i^2}}$$

**Remark.** If  $f(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n x_i$  and  $A = [0, 1]$ , then  $c_i = \frac{1}{n}$  and the bound recovers the *Hoeffding inequality* as  $e^{-2nt^2}$ .



## Chapter 3

# Rademacher Complexity

Preliminary on the relation between  $L_S(h)$  and  $L(h)$  (empirical risk and population risk):

For a fixed  $h$  (independent of  $S$ ),  $\mathbb{E}_S [L_S(h)] = L(h)$ .

By Hoeffding: w.p.  $\geq 1 - \delta$ :  $L(h) - L_S(h) \leq \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$ .

### 3.1 Uniform Convergence

Motivation: we want to study  $L(\hat{h}_{ERM})$  and compare it against  $h^* \in \operatorname{argmin}_{h \in \mathcal{H}} L(h)$ . We want to bound the difference  $L(\hat{h}_{ERM}) - L(h^*)$ , which is also referred to as the “**excess risk**”.

$$L(\hat{h}_{ERM}) - L(h^*) = \left( L(\hat{h}_{ERM}) - L_S(\hat{h}_{ERM}) \right) + \left( L_S(\hat{h}_{ERM}) - L_S(h^*) \right) + (L_S(h^*) - L(h^*))$$

where the second term is smaller or equal to 0 by definition, and the third term can be bounded using the Hoeffding inequality as  $h^*$  does not depend on  $S$ .

Consequently, our aim becomes bounding the first term and we define the following **generalization gap**:

**Definition 3.1.1** (Uniform Convergence).

$$L(\hat{h}_{ERM}) - L_S(\hat{h}_{ERM}) \leq \sup_{h \in \mathcal{H}} (L(h) - L_S(h))$$

, where the bounded difference is called the generalization gap.

**Theorem 3.1.1** (Generalization Bound for finite hypothesis class). If  $\mathcal{H}$  is finite, then for any  $\delta \in (0, 1)$ , we have

$$\text{w.p. } \geq 1 - \delta, \sup_{h \in \mathcal{H}} (L(h) - L_S(h)) \leq \sqrt{\frac{\log |\mathcal{H}| + \log \frac{1}{\delta}}{2n}}$$

**Remark.** If  $n \gg \log |\mathcal{H}|$ , excess risk  $\rightarrow 0$ .

What if  $\mathcal{H}$  is infinite?

– Idea: Reduce infinite case to finite case.

## 3.2 Rademacher Complexity

**Notation:** Given  $\mathcal{H}$  and  $\ell$ , define the family of loss mappings:

$$\mathcal{G} = \{g_h : (x, y) \mapsto \ell(h(x), y), h \in \mathcal{H}\}$$

where  $z = (x, y) \sim P$ ,  $z_i = (x_i, y_i)$ ,  $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ , and  $L(h) = \mathbb{E}_{z \sim P} [g_h(z)]$ ,  $L_S(h) = \frac{1}{n} \sum_{i=1}^n g_h(z_i)$ .

$$\sup_{h \in \mathcal{H}} (L(h) - L_S(h)) = \sup_{g \in \mathcal{G}} \left( \mathbb{E}_{z \sim P} [g(z)] - \frac{1}{n} \sum_{i=1}^n g(z_i) \right)$$

**Definition 3.2.1 (Empirical Rademacher Complexity).** Let  $\mathcal{G}$  be a set of functions mapping  $\mathcal{Z} \rightarrow \mathbb{R}$ . Let  $S = \{z_1, \dots, z_n\} \subseteq \mathcal{Z}$ .

The **empirical Rademacher complexity** of  $\mathcal{G}$  with respect to the simple set  $S$  is:

$$R_S(\mathcal{G}) = \mathbb{E}_{\sigma_1, \dots, \sigma_n} \left[ \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \sigma_i g(z_i) \right]$$

where  $\sigma_i = \begin{cases} +1 & \text{w.p. } \frac{1}{2} \\ -1 & \text{w.p. } \frac{1}{2} \end{cases}$  i.i.d (called Rademacher random variables).

**Remark.** Rademacher complexity measures the ability of a function class to fit random noise

$$R_S(\mathcal{G}) = \mathbb{E}_{\vec{\sigma}} \left[ \sup_{g \in \mathcal{G}} \frac{1}{n} \langle \vec{\sigma}, \vec{g}_s \rangle \right]$$

**Definition 3.2.2 (Rademacher Complexity).** Let  $P$  be a distribution over  $\mathcal{Z}$ .

For an integer  $n \geq 1$ , the **Rademacher complexity** of  $\mathcal{G}$  is

$$R_n(\mathcal{G}) = \mathbb{E}_{S \sim P^n} [R_S(\mathcal{G})]$$

**Theorem 3.2.1 (Generalization Bound using Rademacher Complexity).** Let  $\mathcal{G}$  be a function class mapping  $\mathcal{Z}$  to  $[0, 1]$ ,  $S = \{z_1, \dots, z_n\} \sim P^n$ . Then for any  $\delta \in (0, 1)$ :

$$\text{w.p.} \geq 1 - \delta, \sup_{g \in \mathcal{G}} \left( \mathbb{E}_{z \sim P} [g(z)] - \frac{1}{n} \sum_{i=1}^n g(z_i) \right) \leq 2R_n(\mathcal{G}) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

$$\text{w.p.} \geq 1 - \delta, \sup_{g \in \mathcal{G}} \left( \mathbb{E}_{z \sim P} [g(z)] - \frac{1}{n} \sum_{i=1}^n g(z_i) \right) \leq 2R_S(\mathcal{G}) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2n}}$$

**Proof.** Define  $f(z_1, \dots, z_n) = \sup_{g \in \mathcal{G}} (\mathbb{E}_{z \sim P} [g(z)] - \frac{1}{n} \sum_{i=1}^n g(z_i))$ .

Consider ■