

Machine Learning Theory Notes

Gaotang Li

February 10, 2023

Contents

I	Generalization Theory	2
1	Supervised Learning Framework	3
1.1	Basic Setups	3
1.2	Empirical Risk Minimization	3
1.3	Questions	4
2	Concentration Inequality	5
2.1	Chebyshev's Inequality	5
2.2	Hoeffding Inequality	6
2.3	Bounded Difference Concentration Inequality	7
3	Rademacher Complexity	8
3.1	Uniform Convergence	8
3.2	Rademacher Complexity	8
4	VC-Dimension	11
4.1	Growth Function Bounds	11
4.2	More on VC-Dimension	13
5	Margin Theory	16
5.1	Basic Setups	16

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 μ and finite non-zero variance σ^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 λ' is between 0 and λ .

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

3.1 Uniform Convergence

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

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

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).

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

, 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 ℓ , 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. Step 1: Relate the sup terms to the expectation of sups using Mcdiarmid's ineq

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

Consider $\{z_1, \dots, z_n\}$ and $\{z'_1, \dots, z'_n\}$ that only differs by 1 point (i.e. $z_k \neq z'_k, z_i = z'_i \forall i \neq k$).

$$\begin{aligned} f(z_1, \dots, z_n) &= \sup_{g \in \mathcal{G}} \left(\mathbb{E} [g(z)] - \frac{1}{n} \sum_{i=1}^n g(z'_i) + \frac{1}{n} \sum_{i=1}^n g(z'_i) - \frac{1}{n} \sum_{i=1}^n g(z_i) \right) \\ &\leq \sup_{g \in \mathcal{G}} \left(\mathbb{E} [g(z)] - \frac{1}{n} \sum_{i=1}^n g(z'_i) \right) + \sup_{g \in \mathcal{G}} \left(\frac{1}{n} \sum_{i=1}^n g(z'_i) - \frac{1}{n} \sum_{i=1}^n g(z_i) \right) \\ &= f(z'_1, \dots, z'_n) + \sup_{g \in \mathcal{G}} \left(\frac{1}{n} g(z'_k) - \frac{1}{n} g(z_k) \right) \\ &\leq f(z'_1, \dots, z'_n) + \frac{1}{n} \end{aligned}$$

Similarly, $f(z'_1, \dots, z'_n) - f(z_1, \dots, z_n) \leq \frac{1}{n}$. Combining them we can get that $|f(z_1, \dots, z_n) - f(z'_1, \dots, z'_n)| \leq \frac{1}{n}$.

Applying the Mcdiarmid's inequality, we can get the following bound:

$$\text{w.p.} \geq 1 - \delta, f(z_1, \dots, z_n) - \mathbb{E}[f(z_1, \dots, z_n)] \leq \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

Step 2: Bound $\mathbb{E}_S [\sup_{g \in \mathcal{G}} (\mathbb{E}_{z \sim P} [g(z)] - \frac{1}{n} \sum_{i=1}^n g(z_i))]$ by Rademacher Complexity

Draw a fresh set of n samples $S' = \{z'_1, \dots, z'_n\} \sim P^n$. Fix S , we have

$$\begin{aligned} \sup_{g \in \mathcal{G}} \left(\mathbb{E}_{z \sim P} [g(z)] - \frac{1}{n} \sum_{i=1}^n g(z_i) \right) &= \sup_{g \in \mathcal{G}} \left(\mathbb{E}_{S'} \left[\frac{1}{n} \sum_{i=1}^n g(z_i) \right] - \frac{1}{n} \sum_{i=1}^n g(z_i) \right) \\ &= \sup_{g \in \mathcal{G}} \left(\mathbb{E}_{S'} \left[\frac{1}{n} \sum_{i=1}^n (g(z'_i) - g(z_i)) \right] \right) \\ &\leq \mathbb{E}_{S'} \left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n (g(z'_i) - g(z_i)) \right] \end{aligned}$$

Taking expectation over S on both sides generate that

$$\begin{aligned} \mathbb{E}_S \left[\sup_{g \in \mathcal{G}} \left(\mathbb{E}_{z \sim P} [g(z)] - \frac{1}{n} \sum_{i=1}^n g(z_i) \right) \right] &\leq \mathbb{E}_{S, S'} \left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n (g(z'_i) - g(z_i)) \right] \\ &= \mathbb{E}_{S, S', \sigma} \left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \sigma_i (g(z'_i) - g(z_i)) \right] \quad ? \\ &\leq \mathbb{E}_{S, S', \sigma} \left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \sigma_i g(z'_i) \right] + \mathbb{E}_{S, S', \sigma} \left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n -\sigma_i g(z_i) \right] \\ &= 2R_n(\mathcal{G}) \end{aligned}$$

Combining the result from **step 1** and **step 2**, we prove the first inequality in the theorem.

Step 3: Prove $R_n(\mathcal{G})$ and $R_S(\mathcal{G})$ are close Similar to step 1, we can verify that $R_S(\mathcal{G})$ satisfies the bounded difference property.

Apply Mcdiarmid's inequality, we can get that

$$\text{w.p.} \geq 1 - \delta, R_n(\mathcal{G}) \leq R_S(\mathcal{G}) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

Combining the outputs from step 1 - 3 and replacing δ with $\delta/2$ gives the second inequality. ■

Chapter 4

VC-Dimension

In this chapter, we only consider the binary classification case with the 0-1 loss, *i.e.* $y = \{\pm 1\}$ and $\mathcal{G} = \{(x, y) \mapsto \mathbb{1}[h(x) \neq y] : h \in \mathcal{H}\}$.

4.1 Growth Function Bounds

Lemma 4.1.1. $R_n(\mathcal{G}) = \frac{1}{2} R_n(\mathcal{H})$

Proof. Given $S = \{(x_i, y_i)\}_{i=1}^n$, we have

$$\begin{aligned} R_S(\mathcal{G}) &= \mathbb{E}_{\vec{\sigma}} \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i \mathbb{1}[h(x_i) \neq y_i] \right] \\ &= \mathbb{E}_{\vec{\sigma}} \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i \frac{1 - y_i h(x_i)}{2} \right] \\ &= \frac{1}{2} \mathbb{E}_{\vec{\sigma}} \left[\frac{1}{n} \sum_{i=1}^n \sigma_i + \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i (-y_i) h(x_i) \right] \\ &= \frac{1}{2} \mathbb{E}_{\vec{\sigma}} \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i h(x_i) \right] \\ &= \frac{1}{2} R_S(\mathcal{H}) \end{aligned}$$

■

Remark. It then becomes natural to bound $R_n(\mathcal{H})$.

Definition 4.1.1 (Growth Function). The growth function $\Pi_{\mathcal{H}}: \mathbb{N} \rightarrow \mathbb{N}$ for a hypothesis class \mathcal{H} that maps to $y = \{\pm 1\}$ is defined as

$$\Pi_{\mathcal{H}}(n) = \sup_{x_1, \dots, x_n \in \mathcal{X}} |\{(h(x_1), \dots, h(x_n)) : h \in \mathcal{H}\}|$$

Remark. This definition defines the set of all possible predictions on a given set of inputs.

Theorem 4.1.1 (Generalization bound using VC-dimension). Let \mathcal{H} be a hypothesis class taking values $y = \{\pm 1\}$. Then

$$R_n(\mathcal{H}) \leq \sqrt{\frac{2 \log \Pi_{\mathcal{H}}(n)}{n}}$$

Proof. Let $S = \{x_1, \dots, x_n\}$, $Q = Q_S = \{(h(x_1), \dots, h(x_n)) : h \in \mathcal{H}\}$.

We want to show that $R_S(\mathcal{H}) \leq \sqrt{\frac{2 \log |Q|}{n}}$

$$\begin{aligned} R_S(\mathcal{H}) &= \mathbb{E}_{\vec{\sigma}} \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i h(x_i) \right] \\ &= \mathbb{E}_{\vec{\sigma}} \left[\sup_{\vec{v} \in Q} \frac{1}{n} \sum_{i=1}^n \sigma_i v_i \right] \end{aligned} \quad \text{Apply Hoeffding}$$

Then for all $\lambda > 0$,

$$\begin{aligned} e^{\lambda R_S(\mathcal{H})} &= e^{\lambda \mathbb{E}_{\vec{\sigma}} \left[\sup_{\vec{v} \in Q} \frac{1}{n} \sum_{i=1}^n \sigma_i v_i \right]} \\ &\leq \mathbb{E}_{\vec{\sigma}} \left[e^{\lambda \sup_{\vec{v} \in Q} \frac{1}{n} \sum_{i=1}^n \sigma_i v_i} \right] && \text{Jensen's ineq} \\ &\leq \mathbb{E}_{\vec{\sigma}} \left[\sum_{\vec{v} \in Q} e^{\lambda \frac{1}{n} \sum_{i=1}^n \sigma_i v_i} \right] && \text{why?} \\ &= \sum_{\vec{v} \in Q} \mathbb{E}_{\vec{\sigma}} \left[e^{\lambda \frac{1}{n} \sum_{i=1}^n \sigma_i v_i} \right] \\ &\leq \sum_{\vec{v} \in Q} e^{\frac{\lambda^2}{2n}} && \text{by Hoeffding} \\ &= |Q| e^{\frac{\lambda^2}{2n}} \end{aligned}$$

This gives that $R_S(\mathcal{H}) \leq \frac{1}{\lambda} \log |Q| + \frac{\lambda}{2n}$

Choose $\lambda = \sqrt{2n \log |Q|}$ and we can get that $R_S(\mathcal{H}) \leq \sqrt{\frac{2 \log |Q|}{n}}$ ■

Remark. Discussions about the growth function:

- When \mathcal{H} is finite, we have that $\Pi_{\mathcal{H}}(n) \leq |\mathcal{H}|$

$$R_n(\mathcal{H}) \leq \sqrt{\frac{2 \log |\mathcal{H}|}{n}} \text{ recovers Thm 1}$$

- When \mathcal{H} is “super power”, $\Pi_{\mathcal{H}}(n) = 2^n$, *i.e. overfitting*.

$$R_n(\mathcal{H}) \leq \sqrt{\frac{2 \log 2^n}{n}} = \sqrt{2 \log 2}$$

- What if the growth function is in-between, a polynomial function?

Suppose $\Pi_{\mathcal{H}}(n) \leq n^d$, we have that

$$R_n(\mathcal{H}) \leq \sqrt{\frac{2d \log n}{n}} \rightarrow 0 \text{ if } n \gg d \log d$$

Definition 4.1.2 (VC-dimension). The VC-dimension of a class of hypothesis function \mathcal{H} is

$$\text{VC}(\mathcal{H}) = \max\{n : \Pi_{\mathcal{H}}(n) = 2^n\}$$

Definition 4.1.3 (Shatter). $S = \{x_1, \dots, x_n\}$ can be shattered by \mathcal{H} if $\forall y_1, \dots, y_n \in \{\pm 1\}, \exists h \in \mathcal{H}$ s.t. $h(x_i) = y_i$ for all $i = \{1, \dots, n\}$.

Remark. The VC-dimension is the maximum size of a sample set S that can be **shattered** by \mathcal{H} .

Example (Threshold Function). Let $\mathcal{X} = \mathbb{R}, \mathcal{H} = \{h_a : a \in \mathbb{R}\}, h_a \in \mathcal{H}, h_a(x) = \begin{cases} +1, & \text{if } x \geq a \\ -1, & \text{if } x < a \end{cases}$

Then **VC** – **dim**(\mathcal{H}) = 1

Proof. 1. any input $x \in \mathbb{R}$ can be shattered

$$h_{x-1}(x) = +1, \quad h_{x+1}(x) = -1$$

2. any inputs $x_1, x_2 \in \mathbb{R}$ cannot be shattered

$$x_1 \leq x_2, \text{ impossible to label } (+1, -1)$$

■

Theorem 4.1.2 (growth function bound). Let \mathcal{H} be a hypothesis class with VC-dimension d . Then,

$$\forall n \gg d: \Pi_{\mathcal{H}}(n) \leq \left(\frac{e^n}{d}\right)^d \leq n^d \text{ if } d \geq 3$$

Theorem 4.1.3 (Generalization Bound Using VC-Dimension). Let \mathcal{H} be a hypothesis class taking values in $y = \{\pm 1\}$ and has VC-dim d . Consider the 0-1 loss.

Then, for all $\delta \in (0, 1)$,

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

Remark. This VC-dimension bound requires $n \gg d$. In other words, it is effective when the hypothesis class is relatively less expressive.

4.2 More on VC-Dimension

First we look at more examples illustrating the concept of VC-dimension.

Example (Axis-aligned rectangles). Let $\mathcal{X} = \mathbb{R}, \mathcal{H} = \{h_{a,b,c,d} : a, b, c, d \in \mathbb{R}\}$, and $h \in \mathcal{H}$ be the form

$$h_{a,b,c,d}(x) = \begin{cases} 1 & \text{if } x_1 \in [a, b], x_2 \in [c, d] \\ -1 & \text{otherwise} \end{cases}$$

Then we have **Vc-dim** \mathcal{H} = 4.

Proof. 1. there exists 4 points that can be shattered **exists or for all?**

2. Any 5 points cannot be shattered

Choose the minimum axis-aligned rectangle that contains all 5 points, then it is impossible to label the sides +1 while labeling inside one -1

Example (Linear Functions). Let $\mathcal{X} = \mathbb{R}$, $\mathcal{H} = \{h_w : w \in \mathbb{R}^d\}$, and

$$h_w(x) = \text{sign}(w^T x) = \begin{cases} 1 & \text{if } w^T x \geq 0 \\ -1 & \text{if } w^T x < 0 \end{cases}$$

Then $\text{Vc-dim}(\mathcal{H}) = d$.

Proof. 1. $\exists d$ points that can be shattered Same Question exists or for all?

Choose $x_1, \dots, x_d \in \mathbb{R}^d$ that are linearly independent.

Then for all $y_1, \dots, y_d \in \{\pm 1\}$, we can find a $w \in \mathbb{R}^d$ such that $w^T x_i = y_i$, for all $i = 1, \dots, d$ by solving the set of linear equations.

2. Any $d + 1$ point cannot be shattered

Assume for the sake of contradiction that there exists $d + 1$ points: x_1, \dots, x_{d+1} that can be shattered.

In formal terms, $\exists \alpha = (\alpha_1, \dots, \alpha_{d+1})$ s.t. $\sum_{i=1}^{d+1} \alpha_i x_i = 0$, $\alpha \neq 0$, i.e. \exists a coordinate $k \in \{1, \dots, d+1\}$ s.t. $\alpha_k \neq 0$. WLOG we can assume $\alpha_k > 0$.

For all $w \in \mathbb{R}^d$, we must have $\sum_{i=1}^{d+1} \alpha_i w^T x_i = 0$. why?

Then let $y_i = \text{sign}(\alpha_i)$, $i = 1, \dots, d+1$. $\exists w \in \mathbb{R}^d$ s.t. $\text{sign}(w^T x_i) = y_i$.

Then we find the contradiction:

$$0 = \sum_{i=1}^{d+1} \alpha_i (w^T x_i) < 0 \quad \text{opposite sign}$$

Example (Sine Function). Let $\mathcal{X} = \mathbb{R}$, $\mathcal{H} = \{h_\omega : \omega \in \mathbb{R}\}$, and $h = \text{sign}(\sin(\omega x))$

Then $\text{Vc-dim}(\mathcal{H}) = \infty$.

Proof. It suffices to show that $\exists n$ points that can be shattered, for any n .

Consider n points, $x_i = 2^{-i}$ ($i = 1, \dots, n$) and any labeling $y_1, \dots, y_n \in \{\pm 1\}$.

Define $\frac{\omega}{\pi} = (y'_n y'_{n-1} \dots y'_1 1)_2$ in terms of binary integer, where $y'_i = \begin{cases} 0 & \text{if } y_i = 1 \\ 1 & \text{if } y_i = -1 \end{cases}$

WTS $\text{sign}(\sin(\omega x_i)) = y_i$,

which can be realized through

$$\frac{\omega x_i}{\pi} = \frac{\omega}{\pi} 2^{-i} = (y'_n y'_{n-1} \dots y'_1 1)_2$$

Not fully understand

Theorem 4.2.1 (VC-dimension in finite precision). Let \mathcal{H} be parametrized by p parameters, with each stored in k bits. $\mathcal{H} = \{h_\theta, \theta \in \mathbb{R}^p\}$, then $\text{VC-dim}(\mathcal{H}) \leq k \cdot p$.

Proof. There are $(2^k)^p$ choices for $\theta = (\theta_1, \dots, \theta_p)$, and then

$$2^{\text{Vc-dim}(\mathcal{H})} \leq |\mathcal{H}| \leq 2^{kp}$$

Remark (Limitation of VC-dimension).

$$\begin{aligned} L(h) - L_S(h) &\leq \tilde{O} \left(\sqrt{\frac{VC - \dim(\mathcal{H})}{n}} \right) \\ &\leq \tilde{O} \left(\sqrt{\frac{\#\text{params}}{n}} \right) \end{aligned}$$

If $\# \text{ params} \gg \# \text{ samples}$, the bound will become vacuous.

Chapter 5

Margin Theory

We focus on the binary classification setting where $y = \{\pm 1\}$.

5.1 Basic Setups

Definition 5.1.1 (Margin). The margin of a function $h: \mathcal{X} \rightarrow \mathbb{R}$ at a point $x \in \mathcal{X}$ labeled with $y \in \{\pm 1\}$ is $yh(x)$.

Remark. We have $\hat{y} = \text{sign}(h(x))$; and a classification is correct when $yh(x) > 0$.

Definition 5.1.2 (Margin Loss). For any $\gamma > 0$, define γ -margin loss as

$$\ell_\gamma(y', y) = \ell_\gamma(yy') = \begin{cases} 1, & \text{if } yy' \leq 0 \\ 1 - \frac{yy'}{\gamma} & \text{if } 0 < yy' < \gamma \\ 0, & \text{if } yy' \geq \gamma \end{cases}$$

Remark. Margin Loss \geq 0-1 loss (in terms of their graphs).

Definition 5.1.3 (Population & Empirical Risk for Margin Loss).

$$L_\gamma(h) = \mathbb{E}_{(x,y) \sim P} [\ell_\gamma(h(x), y)]$$

$$L_{\gamma,S}(h) = \frac{1}{n} \sum_{i=1}^n \ell_\gamma(h(x_i), y_i)$$

Remark. $\ell_\gamma(\cdot)$ is $\frac{1}{\gamma}$ -Lipschitz.

SideNote: We say $f: \mathbb{R} \rightarrow \mathbb{R}$ is C -Lipschitz if $|f(x) - f(x')| \leq C|x - x'|$ for all $x, x' \in \mathbb{R}$. OR equivalently, $|f'(x)| \leq C, \forall x \in \mathbb{R}$.

Lemma 5.1.1 (Talagrand's Lemma). Let $\phi: \mathbb{R} \rightarrow \mathbb{R}$ be a C -Lipschitz function. Then,

$$R_S(\phi \odot \mathcal{H}) \leq C \cdot R_S(\mathcal{H})$$

where $\phi \odot \mathcal{H} = \{z \mapsto \phi(h(z)): h \in \mathcal{H}\}$.