# STAT576
# Empirical Process Theory

Pingbang Hu

March 27, 2024

**Abstract**

This is a graduate-level theoretical statistics course taught by Sabyasachi Chatterjee at University of Illinois Urbana-Champaign, aiming to provide an introduction to empirical process theory with applications to statistical $M$-estimation, non-parametric regression, classification and high dimensional statistics.

While there are no required textbooks, some books do cover (almost all) part of the material in the class, e.g., Van Der Vaart and Wellner's *Weak Convergence and Empirical Processes* [VW96].

This course is taken in Fall 2023, and the date on the cover page is the last updated time.

# Contents

# Chapter 1

# Introduction

## Lecture 1: Introduction to Mathematical Statistics

### 1.1 What is Empirical Process Theory?

21 Aug. 9:00

This subject started in the 1930s with the study of the empirical CDF.

**Definition 1.1.1** (Empirical CDF). Given inputs i.i.d. data points $X_1, \dots, X_n \sim \mathbb{P}$, the *empirical CDF* is

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \le t}.$$

The classical result is that, fixing $t$, $F_n(t) \to F(t)$ almost surely.

**Note.** At the same time, $\sqrt{n}(F_n(t) - F(t)) \to \mathcal{N}(0, F(t)(1 - F(t)))$ in distribution.

On the other hand, we can also ask does this convergence happens if we jointly consider all possible $t \in \mathbb{R}$. By the Glivenko-Cantelli theorem, $\sup_{t \le \mathbb{R}} |F_n(t) - F(t)| \overset{n \to \infty}{\to} 0$ almost surely, so the answer is again yes.

Now, we're ready to see a "canonical" example of an empirical process.

**Example** (Canonical empirical process). The *canonical empirical process* is the family of random variables $\{F_n(t)\}_{t \in \mathbb{R}}$, i.e., a stochastic process.

By considering a general class of functions, we have the following.

**Definition 1.1.2** (Empirical process). Let $\chi$ be the domain, $\mathbb{P}$ be a distribution on $\chi$, and $\mathscr{F}$ be the class of function such that $\chi \to \mathbb{R}$. The *empirical process* is the stochastic process indexed by functions in $\mathscr{F}$, $\{G_n(f) \colon f \in \mathscr{F}\}$ where

$$G_n(f) = \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}\left[f(X)\right]$$

and $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$.

**Remark.** The empirical process is a family of mutually dependent random variables, all of them being functions of the same inherent randomness in the i.i.d. data $X_1, \dots, X_n$.

Now, two questions arise.

### 1.1.1 Uniform Law of Large Numbers

As $n \to 0$, whether

$$S_n(\mathscr{F}) \coloneqq \sup_{f \in \mathscr{F}} |G_n(f)| \to 0,$$

and if so, at what rate?

> **Remark.** The rate of convergence of the law of large numbers uniformly over a class of functions $\mathscr{F}$ determines the performance of many types of statistical estimators as we will see.

We will spend most of this course just on this topic with applications. We will show that $S(\mathscr{F})$ concentrates around its expectation and will bound $\mathbb{E}[S(\mathscr{F})]$.

### 1.1.2 Uniform Central Limit Theorem

The most general probabilistic question one can ask is the following.

> **Problem.** What is the joint distribution of the empirical process?

> **Answer.** For a given sample size, it's most often intractable to be able to calculate the joint distribution exactly. One can then use asymptotics when the sample size $n$ is very large to derive limiting distributions. By the regular central limit theorem, $\sqrt{n}G_n(f) \xrightarrow{d} \mathcal{N}(0, \operatorname{Var}[f(X)])$.for any $f$. We want to understand if this holds uniformly (jointly) over $f \in \mathscr{F}$ in some sense. ⊛

We first motivate this through an example.

> **Example** (Uniform empirical process)**.** Consider
>
> - $X_1, \dots, X_n$ i.i.d. from $\mathcal{U}(0,1)$.[a]
> - $\mathscr{F} = \{\mathbb{1}_{[-\infty,t]} : t \in \mathbb{R}\}$
> - $U_n(t) = \sqrt{n}(F_n(t) - t)$ where $F_n$ is the empirical CDF.
>
> We can view $U_n(t)$ as a collection of random variables one for each $t \in (0,1)$, or just as a random function. Then this stochastic process $\{U_n(t) : t \in (0,1)\}$ is called the *uniform empirical process*.
>
> Then, the CLT states that for each $t \in [0,1]$, $U_n(t) \to \mathcal{N}(0, t - t^2)$ as $n \to \infty$. Moreover, for fixed $t_1, \dots, t_k$, the multivariate CLT implies that $(U_n(t_1), \dots, U_n(t_k)) \xrightarrow{d} \mathcal{N}(0, \mathbf{\Sigma})$ where $\mathbf{\Sigma}_{ij} = \min(t_i, t_j) - t_i t_j$.
>
> [a] $\mathcal{U}$ denotes the uniform distribution.

From this example, one can ask questions like the following.

> **Problem.** Does the entire process $\{U_n(t) : t \in [0,1]\}$ converge in some sense? If so, what is the limiting process?

> **Answer.** The limiting process is an object called the *Brownian Bridge*. This was conjectured by Doob and proved by Donsker. ⊛

Other than that, how do we characterize the convergence of stochastic processes in distribution to another stochastic process? How do we generalize this result for a general function class $\mathscr{F}$ defined on a probability space $\chi$? What are some statistical applications of such process convergence results? This is a classical topic and in the last few weeks of this course, we will touch upon some of these questions.

## 1.2 Applications of Uniform Law of Large Numbers

Next, we see one major example where the uniform law of large numbers can be applied.

### 1.2.1 $M$-Estimators

Consider the class of estimators called "$M$-estimator", which is of the form

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} M_\theta(X_i),$$

where $X_1, \dots, X_n$ taking values in $\chi$, $\Theta$ is the parameter space, and $M_\theta \colon \chi \to \mathbb{R}$ for each $\theta \in \Theta$. Let's see some examples.

**Example** (Maximum log-likelihood). $M_\theta(X) = -\log p_\theta(X)$ for a class of densities $\{p_\theta \colon \theta \in \Theta\}$, then $\hat{\theta}$ is the *Maximum log-likelihood* of $\theta$.

There are lots of examples of "local estimators" as well.

**Example** (Mean). $M_\theta(x) = (x - \theta)^2$.

**Example** (Median). $M_\theta(x) = |x - \theta|$.

**Example** ($\tau$ quantile). $M_\theta(x) = Q_\tau(x - \theta)$ where $Q_\tau(x) = (1 - \tau)x\mathbb{1}_{x<0} + \tau x \mathbb{1}_{x \geq 0}$.

**Example** (Mode). $M_\theta(x) = -\mathbb{1}_{|X-\theta| \leq 1}$.

Now, the target quantity for the estimator $\hat{\theta}$ is

$$\theta_0 = \arg\max_{\theta \in \Theta} \mathbb{E}\left[M_\theta(X_1)\right]$$

where $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$. In the asymptotic framework, the two key questions are the following.

**Problem.** Is $\hat{\theta}$ consistent for $\theta_0$? Does $\hat{\theta}$ converge to $\theta_0$ almost surely or in probability as $n \to \infty$? I.e., is $d(\hat{\theta}, \theta_0) \to 0$ for some metric $d$?

**Problem.** What is the rate of convergence of $d(\hat{\theta}, \theta_0)$? For example is it $O(n^{-1/2})$ or $O(n^{-1/3})$?

To answer these questions, one is led to investigate the closeness of the empirical objective function to the population objective function in some uniform sense. Consider $M_n(\theta) = \frac{1}{n}\sum_{i=1}^n M_\theta(X_i)$ and $M(\theta) = \mathbb{E}\left[M_\theta(X_1)\right]$, then

$$\begin{aligned}
\mathbb{P}(d(\hat{\theta}, \theta_0) > \epsilon) &\leq \mathbb{P}\left( \sup_{\theta \colon d(\theta,\theta_0) > \epsilon} M_n(\theta_0) - M_n(\theta) \geq 0 \right) \\
&= \mathbb{P}\left( \sup_{\theta \colon d(\theta,\theta_0) > \epsilon} (M_n(\theta_0) - M(\theta_0) - [M_n(\theta) - M(\theta)]) \geq \inf_{\theta \colon d(\theta,\theta_0) > \epsilon} (M(\theta) - M(\theta_0)) \right) \\
&\leq \mathbb{P}\left( 2 \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \geq \inf_{\theta \colon d(\theta,\theta_0) > \epsilon} (M(\theta) - M(\theta_0)) \right).
\end{aligned}$$

We see that the left-hand side $2\sup_{\theta\in\Theta}|M_n(\theta) - M(\theta)|$ is just $S(\mathscr{F})$ for $\mathscr{F} = \{f_\theta \colon \theta \in \Theta, f_\theta = M_\theta(\cdot)\}$, while the right-hand side $\inf_{\theta \colon d(\theta,\theta_0)>\epsilon} M(\theta) - M(\theta_0)$ is larger than 0.

**Remark.** The last step could be too loose in some problems.

# Lecture 2: Sub-Gaussian Random Variables and the MGF Trick

23 Aug. 9:00

## 1.3 Bounding Supremum of Empirical Process

Most of this course will focus on how to bound the suprema of the empirical process. Let's define it rigorously.

**Problem 1.3.1** (Bounding supremum of empirical process). Given a domain $\chi$, a probability measure $\mathbb{P}$ on $\chi$, data $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, and a function class $\mathscr{F} \ni f \colon \chi \to \mathbb{R}$. We want to find a (non-asymptotically) bound on

$$S_n(\mathscr{F}) = \sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}\left[f(X)\right] \right|.$$

**Answer.** To do this, broadly speaking, we will go through a route with three basic steps:

(a) $S_n(\mathscr{F})$ "concentrates" around its expectation $\mathbb{E}\left[S_n(\mathscr{F})\right]$.

(b) $\mathbb{E}\left[S_n(\mathscr{F})\right] \leq$ the Rademacher complexity of $\mathscr{F}$ via "symmetrization".

(c) Bounding the Rademacher complexity's expected supremum of a "sub-Gaussian process" by a technique called *chaining*.

⊛

Toward this end, we need some basic and fundamental concentration inequalities that are of wide interest and use.

# Chapter 2

# Concentration Inequalities

As we just saw, to solve Problem 1.3.1, we need some basic tools for concentration inequalities. The most celebrated concentration inequality might be the Gaussian tail, which achieves a quadratic exponential decay. Combining this with the classical central limit theorem, we can expect that as $n \to \infty$, approximately the Gaussian tail bound kicks in.

However, to get a concrete, non-asymptotic bound for $S_n(\mathscr{F})$, we would need more sophisticated tools. Let's start with the basics, i.e., the Gaussian distribution.

## 2.1 Gaussian Distribution

For us, the gold standard for concentration would be the Gaussian distribution. The property of the Gaussian distribution we are interested in is its rapid tail decay as we mentioned:

**Lemma 2.1.1.** For $Z \sim \mathcal{N}(0,1)$,

$$\left(\frac{1}{t}-\frac{1}{t^3}\right)\frac{1}{\sqrt{2\pi}}e^{-t^2/2} \le \mathbb{P}(Z \ge t) \le \frac{1}{t}\cdot\frac{1}{\sqrt{2\pi}}e^{-t^2/2}.$$

**Proof.** We want to show

$$\begin{aligned}&\left(\frac{1}{t}-\frac{1}{t^3}\right)\frac{1}{\sqrt{2\pi}}e^{-t^2/2} \le \int_t^\infty \frac{1}{\sqrt{2\pi}}e^{-x^2/2}\,\mathrm{d}x \le \frac{1}{t}\cdot\frac{1}{\sqrt{2\pi}}e^{-t^2/2}\\ \Leftrightarrow &\left(\frac{1}{t}-\frac{1}{t^3}\right)e^{-t^2/2} \le \int_t^\infty e^{-x^2/2}\,\mathrm{d}x \le \frac{1}{t}\cdot e^{-t^2/2}.\end{aligned}$$

Observe that from integration by part (with $x/x$ introduced),

$$\int_t^\infty \frac{x}{x}\cdot e^{-x^2/2}\,\mathrm{d}x = -\frac{e^{-x^2/2}}{x}\bigg|_t^\infty - \int_t^\infty \frac{e^{-x^2/2}}{x^2}\,\mathrm{d}x = \frac{e^{-t^2/2}}{t} - \int_t^\infty \frac{e^{-x^2/2}}{x^2}\,\mathrm{d}x \le \frac{1}{t}\cdot e^{-t^2/2}$$

since the integrand $e^{-x^2/2}/x^2$ is non-negative, which is the desired upper-bound. For the lower-bound, if we again apply integration by part (with $x/x$ introduced again), then

$$\begin{aligned}\int_t^\infty e^{-x^2/2}\,\mathrm{d}x &= \frac{e^{-t^2/2}}{t} - \int_t^\infty \frac{x}{x}\cdot\frac{e^{-x^2/2}}{x^2}\,\mathrm{d}x\\ &= \frac{e^{-t^2/2}}{t} - \left(-\frac{e^{-x^2/2}}{x^3}\bigg|_t^\infty - \int_t^\infty 3\frac{e^{-x^2/2}}{x^4}\,\mathrm{d}x\right)\\ &= \frac{e^{-t^2/2}}{t} - \frac{e^{-t^2/2}}{t^3} + \int_t^\infty 3\frac{e^{-x^2/2}}{x^4}\,\mathrm{d}x\\ &\ge \left(\frac{1}{t}-\frac{1}{t^3}\right)e^{-t^2/2},\end{aligned}$$

since, again, the integrand $3e^{-x^2/2}/x^4$ is non-negative, so the last term is positive, hence we get the desired lower-bound. ■

**Corollary 2.1.1.** For all $t \geq 1$, we have

$$\mathbb{P}(\mathcal{N}(0, \sigma^2) \geq t) \leq e^{-t^2/2\sigma^2}.$$

Now, as is suggested by CLT, the following question arises.

**Problem.** Does Corollary 2.1.1 hold for sums of independent random variables? That is, given i.i.d. $X_1, \ldots, X_n$ with mean $\mu$ and variance $\sigma^2$, whether for all $t \geq 0$,

$$\mathbb{P}(\sqrt{n}(\overline{X} - \mu) \geq t) \leq e^{-t^2/2\sigma^2}?$$

**Answer.** Just invoking CLT is not enough, we need to handle the error term in the normal approximation. We can show this directly for a class of distributions with fast tail decay. ⊛

To go beyond the Gaussian tail bound, let's start with the moment generating function (MGF) trick.

## 2.2 MGF Trick

The MGF trick is easy to develop, but it gives a foundation for all the concentration inequalities we're going to develop. Hence, although it's short, it's worth to make it a separate section.

### 2.2.1 Markov's Inequality

To start with, the most basic tool to bound tail probabilities is the Markov's inequality.

**Lemma 2.2.1** (Markov's inequality)**.** For a non-negative random variable $X \geq 0$,

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

**Note.** Markov's inequality is valid as soon as $\mathbb{E}[X] < \infty$. That is, it holds even when the second moment does not exist.

**Remark.** The rate of tail decay is slow ($O(1/t)$). For the Gaussian, by Lemma 2.1.1, it's $O(e^{-t^2/2})$.

By the above remark, one might ask the following.

**Problem.** Can we derive faster tail decay bounds in general?

**Answer.** Yes, if we assume more moments exist. If all moments exist and in particular the MGF exists, like for the Gaussian, then we can expect faster tail decay. ⊛

### 2.2.2 Chebyshev Inequality

Continuing the discussion on the previous problem, for example, if we assume the second moment exists, then we can get an $O(1/t^2)$ tail decay by Chebyshev inequality.

**Lemma 2.2.2** (Generalized Chebyshev inequality)**.** Given a random variable $X$,

$$\mathbb{P}(|X - \mu| \geq t) = \mathbb{P}(|X - \mu|^p \geq t^p) \leq \min_{p \geq 1} \frac{\mathbb{E}[|X - \mu|^p]}{t^p}.$$

**Proof.** This is directly implied by the Markov's inequality. ■

**Remark** (Chebyshev Inequality). For $p = 2$, we have the usual form

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\operatorname{Var}[X]}{t^2}$$

**Remark.** All tail bounds are derived using Markov's inequality; the clever part is to apply it to the right random variable. In this sense, every tail bound is just Markov's inequality.

### 2.2.3 Crarmer-Chernoff Method

In the same vein, developed by Cramer and Chernoff, if we now assume the MGF exists and apply Markov's inequality, we get the MGF trick.

**Lemma 2.2.3** (MGF trick (Crarmer-Chernoff method)). Given a random variable $X$,

$$\mathbb{P}(X - \mu \geq t) = \mathbb{P}(e^{\lambda(X-\mu)} \geq e^{\lambda t}) \leq \inf_{\lambda > 0} \frac{\mathbb{E}\left[e^{\lambda(X-\mu)}\right]}{e^{\lambda t}}.$$

We will use the MGF trick rather than the generalized Chebyshev's inequality to derive tail bounds because MGF of a sum of independent random variables decomposes as the product of the MGF's. It is messier to work with the $p^{\text{th}}$ moment of a sum of independent random variables.

## 2.3 Hoeffding's Inequality

### 2.3.1 Sub-Gaussian Random Variables

We will now consider a class of distributions whose MGF is dominated by the MGF of a Gaussian. Then, in a very clean way, the MGF trick will give us Gaussian tail bounds for these distributions.

**Definition 2.3.1** (Sub-Gaussian). Given a random variable $X$ with $\mathbb{E}[X] = 0$, we say $X$ is *sub-Gaussian* with variance factor[a] $\sigma^2$ if for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[e^{\lambda X}\right] \leq e^{\frac{\sigma^2 \lambda^2}{2}}.$$

[a] Also called proxy, sub-Gaussian norm, etc.

**Notation.** We write $\operatorname{Subg}(\sigma^2)$ for a compact representation of the class of sub-Gaussian random variables with variance factor $\sigma^2$.

**Remark.** Observe that if $X \in \operatorname{Subg}(\sigma^2)$:

- $-X \in \operatorname{Subg}(\sigma^2)$;
- $X \in \operatorname{Subg}(t^2)$ if $t^2 > \sigma^2$;
- $cX \in \operatorname{Subg}(c\sigma^2)$.

**Lemma 2.3.1** (Equivalent conditions). Given a random variable $X$ with $\mathbb{E}[X] = 0$, the following are equivalent for absolute constants $c_1, \ldots, c_5 > 0$.

(a) $\mathbb{E}\left[e^{\lambda X}\right] \leq e^{c_1^2 \lambda^2}$ for all $\lambda \in \mathbb{R}$.

(b) $\mathbb{P}(|X| \geq t) \leq 2e^{-t^2/c_2^2}$.

(c) $\left(\mathbb{E}\left[|X|^p\right]\right)^{1/p} \leq c_3\sqrt{p}$.

(d) For all $\lambda$ such that $|\lambda| \le 1/c_4$, $\mathbb{E}\left[e^{\lambda^2 X^2}\right] \le e^{c_4^2\lambda^2}$.

(e) For some $c_5 < \infty$, $\mathbb{E}\left[e^{X^2/c_5^2}\right] \le 2$.

**Proof.** Let's just see the first implication from (a) to (b). Given $X \in \text{Subg}(\sigma)$,

$$\mathbb{P}(X \ge t) \le \inf_{\lambda>0} e^{\lambda^2\sigma^2/2 - \lambda t} \le e^{-\frac{t^2}{2\sigma^2}}$$

where the last inequality follows from minimizing the quadratic function $\lambda^2\sigma^2/2 - \lambda t$ whose minimizer is $\lambda^* = t/\sigma^2$. The same bound holds for the left tail and a union bound gives the two-sided version. For a complete proof, see Lemma A.1.1. ■

Let's see some examples of the sub-Gaussian random variables.

**Example** (Rademacher random variable). $\epsilon = \pm 1$ with probability $1/2$ is a $\text{Subg}(1)$ random variable.

**Proof.** We see that

$$\mathbb{E}\left[e^{\lambda\epsilon}\right] = \frac{1}{2}e^{\lambda} + \frac{1}{2}e^{-\lambda} = \frac{1}{2}\sum_{k=1}^{\infty}\left(\frac{\lambda^k}{k!} + \frac{(-\lambda)^k}{k!}\right) = \sum_{k=1}^{\infty}\frac{\lambda^{2k}}{(2k)!} \le 1 + \sum_{k=1}^{\infty}\frac{(\lambda^2)^k}{2^k k!} = e^{\lambda^2/2}$$

since $(2k)! \ge 2^k \cdot k!$. ⊛

In fact, the above can be generalized for any bounded random variable.

**Lemma 2.3.2.** Given $X \in [a, b]$ such that $\mathbb{E}[X] = 0$. Then

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

for all $\lambda \in \mathbb{R}$, i.e., $X \in \text{Subg}((b-a)^2/4)$.

**Proof.** We will prove this with a worse constant. Let $X' \overset{\text{i.i.d.}}{\sim} X$ be an i.i.d. copy, then

$$\mathbb{E}\left[e^{\lambda X}\right] = \mathbb{E}\left[e^{\lambda(X - \mathbb{E}[X'])}\right] = \mathbb{E}\left[e^{\lambda X}\cdot e^{-\lambda(\mathbb{E}[X'])}\right] \le \mathbb{E}\left[e^{\lambda X}\right]\cdot\mathbb{E}\left[e^{-\lambda X'}\right] = \mathbb{E}\left[e^{\lambda(X-X')}\right],$$

where we have used the Jensen's inequality for $e^{-\lambda\mathbb{E}[X']} \le \mathbb{E}\left[e^{-\lambda X'}\right]$.[a] Now we introduce a Rademacher random variable $\epsilon = \pm 1$, to further write

$$\mathbb{E}\left[e^{\lambda X}\right] \le \mathbb{E}_{X,X'}\left[e^{\lambda(X-X')}\right] = \mathbb{E}_{X,X',\epsilon}\left[e^{\lambda\cdot\epsilon(X-X')}\right] = \mathbb{E}_{X,X'}\left[\mathbb{E}_{\epsilon}\left[e^{\lambda\epsilon(X-X')}\right]\right],$$

and $\mathbb{E}_{\epsilon}\left[e^{\lambda\epsilon(X-X')}\right] \le e^{\frac{\lambda^2(X-X')^2}{2}} \le e^{\frac{\lambda^2(b-a)^2}{2}}$, where we used the above known bound on MGF of a Rademacher random variable, hence overall, we get

$$\mathbb{E}\left[e^{\lambda X}\right] \le \mathbb{E}_{X,X'}\left[e^{\frac{\lambda^2(b-a)^2}{2}}\right] = e^{\frac{\lambda^2(b-a)^2}{2}}.$$

For a complete proof, see Lemma A.1.3. ■

[a] This is a trick called symmetrization. A basic example is $\text{Var}[X] = \frac{1}{2}\mathbb{E}\left[(X-X')^2\right]$.

**Note.** If $a = -1$ and $b = 1$, we get back to the earlier example.

Just like independent Gaussians, sums of independent sub-Gaussians remain sub-Gaussian.

**Lemma 2.3.3** (Closed under convolution). Let $X_i$ be independent random variables with $\mathbb{E}[X_i] = \mu_i$,

and $X_i - \mu_i \in \text{Subg}\left(\sigma_i^2\right)$. Then

$$\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i \in \text{Subg}\left(\sum_{i=1}^n \sigma_i^2\right).$$

**Proof.** We simply observe that

$$\mathbb{E}\left[e^{\lambda \sum_i (X_i - \mu_i)}\right] = \prod_{i=1}^n \mathbb{E}\left[e^{\lambda (X_i - \mu_i)}\right] \leq e^{\frac{\lambda^2 (\sum_i \sigma_i^2)}{2}}.$$

■

**Lemma 2.3.4.** Let $X_1, \ldots, X_n \sim \text{Subg}(\sigma_i^2)$, not necessary independent. Then for some absolute constant $c > 0$,

$$\mathbb{E}\left[\max_i |X_i|\right] \leq c\sqrt{\log n} \max_{1 \leq i \leq n} \sigma_i.$$

**Proof.** See Lemma A.1.4 for a proof. ■

### 2.3.2 Hoeffding's Inequality

We can now immediately prove the famous Hoeffding's inequality, which is the main tool in our interest.

**Theorem 2.3.1** (Hoeffding's inequality for sub-Gaussian random variables)**.** Let $X_i$ be independent random variables with $\mathbb{E}[X_i] = \mu_i$, and $X_i - \mu_i \in \text{Subg}(\sigma_i^2)$. Then for all $t \geq 0$,[a]

$$\mathbb{P}\left(\left|\sum_{i=1}^n (X_i - \mu_i)\right| \geq t\right) \leq 2\exp\left(\frac{-t^2}{2\sum_i \sigma_i^2}\right).$$

[a] One-sided version holds without the factor 2.

**Proof.** It's immediate from Lemma 2.3.3 and the equivalent condition (b) in Lemma 2.3.1. ■

## Lecture 3: Sub-Exponential Random Variables

For bounded random variables, we can apply Hoeffding's inequality to obtain the following. 25 Aug. 9:00

**Corollary 2.3.1.** Let $X_i \in [a, b]$ be random variables with mean $\mu_i$,

$$\mathbb{P}\left(\sum_{i=1}^n (X_i - \mu_i) \geq t\right) \leq \exp\left(-\frac{2t^2}{n(b-a)^2}\right).$$

As a consequence, if $X_i$ are i.i.d., then

$$\mathbb{P}(\sqrt{n}(\overline{X} - \mu) \geq t) \leq \exp\left(-\frac{-2t^2}{(b-a)^2}\right).$$

Compare this with Gaussian approximation, we then have

$$\mathbb{P}(\sqrt{n}(\overline{X} - \mu) \geq t) \approx \mathbb{P}(\mathcal{N}(0, \sigma^2) \geq t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right),$$

i.e., $\sigma^2 \sim (b-a)^2/4$.[1]

[1] Actually, $\sigma^2 \leq (b-a)^2/4$ always holds.

**Remark** (Comparison between Hoeffding's bound and Gaussian tail bound)**.** We see that

(a) Hoeffding's inequality can be used for any sample size, but Gaussian approximation can only be used when $n$ is large.

(b) As $\sigma^2 \le (b-a)^2/4$, we see that Gaussian approximation gives a tighter tail bound.

(c) Another way to state this is that from CLT we get the asymptotically valid confidence interval for $\mu$ as

$$\left[\overline{X} \pm \frac{\sigma}{\sqrt{n}} Z_{\alpha/2}\right],$$

while from the Hoffding's inequality, we have (finite sample valid) confidence interval

$$\left[\overline{X} \pm \frac{b-a}{2\sqrt{n}} \sqrt{\log \frac{2}{\alpha}}\right],$$

which is much larger.

The above discussion suggests that if the range is very large compared to the variance, then Hoeffding's inequality may not perform very well. Clearly, such random variables exist. Here are some examples.

**Example.** Suppose

$$\mathbb{P}(X = 0) = 1 - 1/k^2$$
$$\mathbb{P}(X = \pm K) = 1/2k^2$$

with $\mathbb{E}[X] = 0$ and $\mathrm{Var}[X] \le 1$. The range is $2K$, which is very large compared to the variance. This is a case where Hoeffding's inequality would not perform very well, in the sense that the confidence interval based on it would be too wide.

Another example is the following.

**Example.** Let $X_1, \dots, X_n$ be i.i.d. Bernoulli$(\lambda/n)$, where each one of them has range 1, but its variance is at most $\frac{\lambda}{n} \ll 1$. Then a direct application of Hoeffding's inequality gives

$$\mathbb{P}\left(\sum_i X_i - \lambda \ge t\right) \le \exp\left(\frac{-2t^2}{n}\right).$$

This suggests that $\sum_i X_i = O(\sqrt{n})$ whereas we know that in this case the distribution of $\sum_i X_i$ is close to the Poisson$(\lambda)$ and thus should be $O(1)$.

On the other hand, the CLT-inspired bound would give the right order. This points out that we would like to be able to replace the range term with the variance in Hoeffding's inequality. This is what is done in Bernstein's inequality which we will discuss next.

Let's see some non-examples.

**Example** (Not sub-Gaussian)**.** Some examples of random variables that are not sub-Gaussians random variables are Cauchy, exponential, and Poisson random variables.

What about a mixture?

**Problem.** Suppose $Z_1, Z_2 \in \mathrm{Subg}(\sigma^2)$ with mean 0, and consider

$$X = \begin{cases} Z_1, & \text{w.p. } p; \\ Z_2, & \text{w.p. } 1-p. \end{cases}$$

Is this a sub-Gaussian random variable?

## 2.4 Bernstein's Inequality

### 2.4.1 Sub-Exponential Random Variables

The main reason for considering the class of sub-Gaussian random variables is that the MGF is finite and thus the MGF trick works. So if we want to extend the MGF trick, we would like to ask the following:

**Problem.** How fast could the tails of a distribution be so that the MGF is finite?

**Answer.** It turns out that we can allow fatter tails than sub-Gaussian, essentially the PDF can decay no slower than an exponential with a proper exponent. ⊛

Consider the following example.

**Example.** Let $Z^2 \sim \chi^2$, then for all $t \geq 1$, $\mathbb{P}(Z^2 > t) = 2\mathbb{P}(Z \geq \sqrt{t}) \leq 2e^{-t/2}$. It is seen that the rate of decrease of the $\chi^2$ tail probability is slower than that of normal. In fact, the MGF of $\chi^2$ is

$$\mathbb{E}\left[e^{\lambda(Z^2-1)}\right] = \begin{cases} \dfrac{e^{-\lambda}}{\sqrt{1-2\lambda}}, & \text{if } 0 \leq \lambda < 1/2; \\ \infty, & \text{if } \lambda \geq 1/2, \end{cases}$$

where we see that the MGF exists in a neighborhood around 0, but not everywhere.

This motivates the following definition.

**Definition 2.4.1** (Sub-exponential). A random variable $X$ is *sub-exponential* with parameters $(\sigma^2, \alpha)$ with mean $\lambda$ if for all $|\lambda| < 1/\alpha$

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

It's then immediate to see that $\text{SubExp}(\sigma^2, \alpha)$ random variables have the same bound on their MGF as a $\text{Subg}(\sigma^2)$ but only for $\lambda$ in the interval $(-\frac{1}{\alpha}, \frac{1}{\alpha})$.

**Example.** For the $\chi^2$ random variable $Z^2$, we have $Z^2 \in \text{SubExp}(2, 4)$.

**Proof.** This is immediate from Definition 2.4.1 since For all $|\lambda| < 1/4$, we have

$$\frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \leq e^{2\lambda^2}.$$

⊛

With Definition 2.4.1, we can extend the MGF trick naturally.

**Lemma 2.4.1** (Tail decay for sub-exponential random variable). Let $X \in \text{SubExp}(\sigma^2, \alpha)$ with mean $\mu$. Then

$$\mathbb{P}(X - \mu \geq t) \leq \begin{cases} e^{-\frac{t^2}{2\sigma^2}}, & \text{if } 0 \leq t \leq \dfrac{\sigma^2}{\alpha}; \\ e^{-\frac{t}{2\alpha}}, & \text{if } t > \dfrac{\sigma^2}{\alpha}. \end{cases}$$

**Proof.** We see that

$$\mathbb{P}(X - \mu \geq t) \leq \inf_{0 \leq \lambda < 1/\alpha} \frac{\mathbb{E}\left[e^{\lambda(X-\mu)}\right]}{e^{\lambda t}} \leq \inf_{0 \leq \lambda < 1/\alpha} e^{\frac{\lambda^2\sigma^2}{2} - \lambda t}.$$

Now, we just need to minimize the exponent, which is a convex quadratic function, in the range $(0, \frac{1}{\alpha})$. The infimum depends on the value of $\alpha$:

- $\frac{t}{\sigma^2} < \frac{1}{\alpha}$: we get the Gaussian bound.

- $\frac{t}{\sigma^2} \geq \frac{1}{\alpha}$: the minimizer is $1/\alpha$, and we get the exponential bound.

■

**Corollary 2.4.1.** Let $X \in \text{SubExp}(\sigma^2, \alpha)$ with mean $\mu$. Then for all $t \geq 0$,

$$\mathbb{P}(|X - \mu| \geq t) \leq 2\exp\left(-\frac{t^2}{2(\sigma^2 + t\alpha)}\right).$$

**Proof.** We see that

$$\mathbb{P}(|X - \mu| \geq t) \leq 2\exp\left(-\min\left\{\frac{t^2}{2\sigma^2}, \frac{t}{2\alpha}\right\}\right) \leq 2\exp\left(-\frac{t^2}{2(\sigma^2 + t\alpha)}\right)$$

by observing $\min(1/u, 1/v) \geq 1/(u+v)$. ■

Just like Lemma 2.3.3 for sub-Gaussian random variables, sub-exponential random variables are also closed under convolution.

**Lemma 2.4.2** (Closed under convolution)**.** Let $X_i \in \text{SubExp}(\sigma_i^2, \alpha_i)$ be all independent with mean $\mu_i$, then

$$\sum_i (X_i - \mu_i) \in \text{SubExp}\left(\sum_i \sigma_i^2, \|\alpha\|_\infty\right).$$

**Proof.** Since

$$\mathbb{E}\left[e^{\lambda \sum_i (X_i - \mu_i)}\right] = \prod_{i=1}^n \mathbb{E}\left[e^{\lambda(X_i - \mu_i)}\right] \leq \prod_{i=1}^n e^{\lambda^2 \sigma_i^2/2} = e^{\lambda^2 \sum_i \sigma_i^2/2}$$

where the inequality holds if $|\lambda| < 1/\alpha_i$ for all $i$, i.e., $|\lambda| < 1/\|\alpha\|_\infty$. ■

### 2.4.2 Bernstein's Inequality

We are now ready to state the generalization of Hoeffding's inequality to sums of independent sub-exponential random variables.

**Theorem 2.4.1** (Bernstein's inequality for sub-exponential random variables)**.** Let $X_i \sim \text{SubExp}(\sigma_i^2, \alpha_i)$ be all independent with mean $\mu_i$, then

$$\mathbb{P}\left(\left|\sum_{i=1}^n (X_i - \mu_i)\right| \geq t\right) \leq 2\exp\left(-\min\left\{\frac{t^2}{2\sum_i \sigma_i^2}, \frac{t}{2\|\alpha\|_\infty}\right\}\right).$$

**Proof.** This is immediate from Lemma 2.4.1 and Lemma 2.4.2. ■

We can restate Bernstein's inequality in a convenient way.

**Corollary 2.4.2.** Let $X_i \sim \text{SubExp}(\sigma_i^2, \alpha_i)$ be all independent with mean $\mu_i$, and let $k \geq \sigma_i, \alpha_i$ for all $i$. Then for all $a_i \in \mathbb{R}$, we have

$$\mathbb{P}\left(\left|\sum_{i=1}^n a_i(X_i - \mu_i)\right| \geq t\right) \leq 2\exp\left(-\min\left\{\frac{t^2}{k^2\|a\|^2}, \frac{t}{k\|a\|_\infty}\right\}\right).$$

**Note.** If we let $a_i = 1/\sqrt{n}$, we obtain an absolute constant $c$ (depending on $k$ only)

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - \mu_i)\right| \geq t\right) \leq \begin{cases} 2e^{-ct^2}, & \text{if } 0 < t < c\sqrt{n}; \\ 2e^{-t\sqrt{n}}, & \text{if } t > c\sqrt{n}. \end{cases}$$

**Remark.** Bernstein's inequality gives the sub-Gaussian tail decay expected from CLT for most $t$. Only in the very rare event regime, does the slower exponential tail decay come in.

# Lecture 4: McDiarmid's Inequality

28 Aug. 9:00

## 2.5 Bounded Difference Concentration Inequality

### 2.5.1 Applications of Berstein's Inequality to Bounded Random Variables

Now we see some applications of Bernstein's inequality, addressing weaknesses of Hoeffding's inequality.

**Lemma 2.5.1.** Let $|X-\mu| \le b$ and $X-\mu$ is Subg$(b^2)$. It's also true that $X-\mu \in \text{SubExp}(2\sigma^2, 2b)$ where $\text{Var}[X] = \sigma^2$.

**Proof.** From $(X-\mu)^k \le (X-\mu)^2|X-\mu|^{k-2} \le (X-\mu)^2 b^{k-2}$, we have

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] = 1 + \frac{\lambda^2}{2}\sigma^2 + \sum_{k=3}^{\infty} \lambda^k \frac{\mathbb{E}[X-\mu]^k}{k!} \le 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda\sigma^2}{2}\sum_{k=3}^{\infty}(|\lambda|b)^{k-2}.$$

The last sum is a geometric series, which converges if $|\lambda| < 1/b$ to

$$1 + \frac{\lambda^2\sigma^2}{2}\left(\frac{1}{1-b|\lambda|}\right).$$

Then from $1 + x \le e^x$, we see that for $|\lambda| < 1/2b$,

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \le e^{\frac{\lambda^2\sigma^2}{2(1-b|\lambda|)}} \le e^{\lambda^2\sigma^2}.$$

■

From this, by directly applying Bernstein's inequality, we have the following.

**Corollary 2.5.1.** Let $X$ be a random variable such that $|X-\mu| \le b$. For any $t > 0$,

$$\mathbb{P}(|X-\mu| \ge t) \le 2\exp\left(\frac{-t^2}{2(2\sigma^2 + t\cdot 2b)}\right).$$

Furthermore, let $X_1, \dots, X_n$ be independent random variables with $\mathbb{E}[X_i] = \mu_i$ and $\text{Var}[X_i] = \sigma_i^2$ such that $|X_i - \mu_i| \le b$ for all $i$. Then for any $t > 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n}(X_i - \mu_i)\right| \ge t\right) \le 2\exp\left(\frac{-t^2}{4\left(\sum_i \sigma_i^2 + tb\right)}\right).$$

In particular, if $\mu_i = \mu$ for all $i$, then

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \ge t\right) \le 2\exp\left(-\frac{nt^2}{4(\sigma^2 + tb)}\right).$$

**Remark.** Observe that in the last line of the proof of Lemma 2.5.1, the inequality is quite loose. This means that we can explicitly maximize the quantity in the exponent over $|\lambda| \in (0, 1/2b)$ to get a higher bound and hence, a better variance factor. This leads to a tighter version of Corollary 2.5.1.

**Corollary 2.5.2.** Let $X_1, \dots, X_n$ be independent random variables with $\mathbb{E}[X_i] = \mu$ and $\text{Var}[X_i] = \sigma^2$

such that $|X_i - \mu| \le b$ for all $i$. Then for any $t > 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i - \mu\right| \ge t\right) \le 2\exp\left(\frac{-t^2/2}{n\sigma^2 + bt/3}\right).$$

In particular,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \ge t\right) \le 2\exp\left(\frac{-nt^2/2}{\sigma^2 + bt/3}\right).$$

**Note.** From Corollary 2.5.2:

- if $t \le 3\sigma^2/b$, the tail of the sample mean behaves like a sub-Gaussian tail;
- if $t > 3\sigma^2/b$, the tail of the sample mean behaves like a sub-exponential tail.

**Remark.** In practice, since we know that the sample mean is $\sqrt{n}$-consistent, we generally look at a sequence of quantiles of the sample mean that is of $O(n^{-1/2})$. Therefore, the tail behavior when $t$ gets large, is practically irrelevant.

By choosing the appropriate $t$ in the above tail bound, we can get the following confidence interval for $\mu$.

**Corollary 2.5.3.** Under the assumption of Corollary 2.5.2,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \le \frac{\sigma}{\sqrt{n}}\sqrt{2\log\frac{2}{\alpha}} + \frac{3b}{3n}\log\frac{2}{\alpha}\right) \ge 1 - \alpha$$

**Proof.** Let

$$\alpha = 2\exp\left(\frac{-t^2}{2(V + bt/3)}\right),$$

then

$$t^2 - \frac{2tb}{3}\log\frac{2}{\alpha} - 2V\log\frac{2}{\alpha} = 0.$$

■

In Corollary 2.5.3, we have an $O(1/\sqrt{n})$ term, which is similar to the one derived from Hoeffding's inequality for bounded random variables. In contrary to Hoeffding's bound, we have an additional lower-order term here.

**Remark.** Observe that the higher order term in Corollary 2.5.3 involves the variance, whereas in the case of Hoeffding, it involves the range. Therefore, for random variables with a large range but is highly concentrated around its mean, the Hoeffding confidence interval would be much wider.

The above remark is demonstrated by the following example.

**Example.** Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Ber}(p)$. Suppose we observe $X_i = 0$ for all $i$, then $\hat{p} = \overline{X} = 0$ and the estimate of $\text{Var}[X_1]$ would be $\hat{p}(1-\hat{p}) = 0$.

Hence, if we plug this estimate of variance into the confidence bound from Bernstein, the length of which would be $O(1/n)$. However, in the case of Hoeffding (which works with the range, in this case, 1), the length would be $O(1/\sqrt{n})$.

### 2.5.2 McDiarmid's Inequality

Now we go back to the discussion about empirical process. We do the first step, i.e., we want to show

$$S_n = \sup_{f \in \mathscr{F}} \left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)]\right|$$

"concentrates" when $\mathscr{F}$ is bounded provided that

$$\sup_{x\in\chi, f\in\mathscr{F}} |f(x)| \le B.$$

One simple example of a bounded function class arises in the task of classification.

**Example** (Classification). Consider $f(x)$ corresponds to the class label of an observation with feature value $x$, then the class is bounded.

However, since $S_n$ falls neither into the category of Hoeffding nor Bernstein, we would need a more general concentration inequalities: the McDiarmid's inequality, which considers the $f$ satisfying the bounded-difference property.[2]

**Definition 2.5.1** (Bounded-difference property). Let $X_1, \dots, X_n$ be i.i.d. random variables on $\chi$. We say $f\colon \chi^n \to \mathbb{R}$ satisfies the *bounded-difference property* with parameters $c_i$ if for all $i$,

$$\sup_{x_1,\dots,x_n,x_i'} |f(x_1,\dots,x_n) - f(x_1,\dots,x_i',\dots,x_n)| \le c_i.$$

**Theorem 2.5.1** (McDiarmid's inequality). Let $X_1, \dots, X_n$ be i.i.d. random variables on $\chi$, and let $f\colon \chi^n \to \mathbb{R}$ satisfying the bounded-difference property with parameters $c_1, \dots, c_n$. Then for any $t > 0$,

$$\mathbb{P}(f(X_1,\dots,X_n) - \mathbb{E}\left[f(X_1,\dots,X_n)\right] \ge t) \le \exp\left(\frac{-2t^2}{\sum_i c_i^2}\right).$$

The same bound holds for the left tail.

**Remark.** The qualitative statement for McDiarmid's inequality is that "a random variable depends on the influence of many independent random variables but not too many on any one of them concentrates".

**Proof.** Typically, $\sum_i c_i = O(1)$ concentration will happen if $\sum_i c_i^2 = o(1)$. For example, if each $c_i = O(1/n)$, then concentration happens but not when all $c_i = 0$ except one of them is 1. ⊛

**Remark.** McDiarmid's inequality is a generalization of Hoffding's inequality.

**Proof.** Let

$$f(x_1,\dots,x_n) = \frac{1}{n}(x_1 + \cdots + x_n).$$

When $X_i$'s are independent and $X_i \in [a_i, b_i]$ for all $i$, it's easy to observe that when we change the $i^{\text{th}}$ argument of $f$, the value of $f$ can change at most by $(b_i - a_i)/n$, i.e., McDiarmid's inequality is satisfied with $c_i \coloneqq (b_i - a_i)/n$, plugging in, we get back Hoffding's inequality. ⊛

With McDiarmid's inequality, we can check that the following holds for bounded function classes $\mathscr{F}$:

$$|S_n(x_1,\dots,x_n) - S_n(x_1,\dots,x_i',\dots,x_n)| \le \frac{2B}{n} =: c_i.$$

Then from McDiarmid's inequality, for any $t > 0$,

$$\mathbb{P}(S_n \ge \mathbb{E}\left[S_n\right] + t) \le \exp\left(\frac{-nt^2}{2B^2}\right) =: \delta,$$

or equivalently, $S_n \le \mathbb{E}\left[S_n\right] + B\sqrt{\frac{2}{n}\log\frac{1}{\delta}}$ with probability at least $1 - \delta$.

[2] Hence it's also known as the *bounded difference inequality*.

**Note.** $B\sqrt{\frac{2}{n}\log\frac{1}{\delta}}$ is a lower order term, i.e., $\mathbb{E}[S_n]$ dominates it.

**Proof.** Since[a]

$$O(B) \geq \mathbb{E}[S_n] \geq \mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} f(x_i) - \mathbb{E}[f(X)]\right|\right] = O\left(\sqrt{\frac{\mathrm{Var}[f(X_1)]}{n}}\right) \approx O\left(\frac{1}{\sqrt{n}}\right).$$

⊛

[a] This upper-bound is pretty weak, and we will eventually work on getting better bounds.

All these imply that *it's enough to bound* $\mathbb{E}[S_n]$.

# Lecture 5: Proof of McDiarmid's Inequality

1 Sep. 9:00

We should note that the usual proof of McDiarmid inequality involves martingale decomposition and Azuma-Hoeffding inequality, a generalization of Hoffding's inequality for martingale difference sequence. However, we will not go with this route; instead, we prove something weaker but tricker.[3]

**Note.** The condition $\sup_{x_1,\dots,x_n,x_i'} |f(x_1,\dots,x_n) - f(x_1,\dots,x_i',\dots,x_n)| \leq c_i$ is equivalent to

$$|f(x_1,\dots,x_n) - f(z_1,\dots,z_n)| \leq \sum_{i=1}^{n} c_i \mathbb{1}_{x_i \neq z_i}.$$

Now, we need one last lemma to prove McDiarmid inequality.

**Lemma 2.5.2.** For all $x \neq y \in \mathbb{R}$,

$$\frac{e^x - e^y}{x-y} \leq \frac{e^x + e^y}{2} \Rightarrow |e^x - e^y| \leq |x-y|\left(\frac{e^x+e^y}{2}\right).$$

**Proof.** Since

$$\frac{e^x - e^y}{x-y} = \int_0^1 e^{sx+(1-s)y}\,\mathrm{d}s = \frac{1}{x-y}\int_x^y e^t\,\mathrm{d}t$$

where we let $t = sx + (1-s)y$. On the other hand, due to convexity, we also have

$$\frac{e^x - e^y}{x-y} = \int_0^1 e^{sx+(1-s)y}\,\mathrm{d}s \leq \int_0^1 s\cdot e^x + (1-s)e^y\,\mathrm{d}s = \frac{e^x+e^y}{2}.$$

■

We're now ready.

**Proof of Theorem 2.5.1.** Firstly, we note that it's equivalent to show that $f(X_1,\dots,X_n) - \mathbb{E}[f] \in \mathrm{Subg}(\sum_i c_i^2/4)$. Without loss of generality, let $\mathbb{E}[f] = 0$, and we want to show that

$$\mathbb{E}\left[e^{\lambda(f(X)-\mathbb{E}[f])}\right] \leq e^{\frac{\lambda^2\sum_i c_i}{8}} \Leftrightarrow M(\lambda) = \mathbb{E}\left[e^{\lambda f(X)}\right] \leq \exp\left(\frac{\lambda^2\left(\sum_i c_i^2\right)}{8}\right) \Leftrightarrow \log M(\lambda) \leq \lambda^2\frac{\sum_i c_i^2}{8}.$$

Observe that since both sides of the inequality is 0 at $\lambda = 0$, it's enough to show

$$\frac{\mathrm{d}\log M(\lambda)}{\mathrm{d}\lambda} = \frac{M'(\lambda)}{M(\lambda)} \leq \lambda\cdot\frac{\sum_i c_i^2}{4}$$

Let $X = (X_1,\dots,X_n)$, and $X' \overset{\text{i.i.d.}}{\sim} X$ be the i.i.d. copy of $X$. Then define the following.

[3] In fact, what we're going to prove is not even a weaker version: we prove something weaker while we really need the original (stronger) statement to hold.

**Notation.** $X^{(i)} \coloneqq (X_1', \dots, X_i', X_{i+1}, \dots, X_n)$ and $X^{[i]} \coloneqq (X_1, \dots, X_{i-1}, X_i', X_{i+1}, \dots, X_n)$.

Note that this implies $X^{(0)} = X$ and $X^{(n)} = X'$. Then, we can show that

$$\begin{aligned} M'(\lambda) &= \mathbb{E}\left[f(X)e^{\lambda f(X)}\right] && \text{As } \mathbb{E}[f] = 0 \text{ and } X, X' \text{ are independent} \\ &= \mathbb{E}\left[\big(f(X) - f(X')\big)e^{\lambda f(X)}\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{n} (f(X^{(i-1)}) - f(X^{(i)})) \cdot e^{\lambda f(X)}\right] \end{aligned}$$

if $i^{\text{th}}$ position of $X$ and $X'$ are swapped, then for the new data $X^{(i-1)}$ and $X^{(i)}$ will also be swapped,

$$\begin{aligned} &= \mathbb{E}\left[\frac{1}{2}\sum_{i=1}^{n}\Big(f(X^{(i-1)}) - f(X^{(i)})\Big) \cdot \Big(e^{\lambda f(X)} - e^{\lambda f(X^{[i]})}\Big)\right] \\ &\le \mathbb{E}\left[\frac{\lambda}{2}\sum_{i=1}^{n}\Big|f(X^{(i-1)}) - f(X^{(i)})\Big| \cdot \Big|f(X) - f(X^{[i]})\Big| \cdot \left(\frac{e^{\lambda f(X)} + e^{\lambda f(X^{[i]})}}{2}\right)\right] \\ &\qquad\qquad \text{from Lemma 2.5.2} \\ &\le \frac{\lambda}{2}\left(\sum_{i=1}^{n} c_i^2\right) \cdot M(\lambda). \end{aligned}$$

■

We note the following.

**Note.** The above proof doesn't even show a weaker version of McDiarmid's inequality.

**Proof.** While in the proof, we need to show

$$\frac{\mathrm{d}\log M(\lambda)}{\mathrm{d}\lambda} = \frac{M'(\lambda)}{M(\lambda)} \le \lambda \cdot \frac{\sum_i c_i^2}{4},$$

we only show

$$\frac{\mathrm{d}\log M(\lambda)}{\mathrm{d}\lambda} = \frac{M'(\lambda)}{M(\lambda)} \le \lambda \cdot \frac{\sum_i c_i^2}{2}.$$

For a complete proof, see Theorem A.1.2. ⊛

### 2.5.3 Applications of McDiarmid's Inequality

**$U$-Statistics**

Let $g\colon \mathbb{R}^2 \to \mathbb{R}$ be a symmetric function, and let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$. Consider

$$U(X) = \frac{1}{\binom{n}{2}}\sum_{j<k} g(X_j, X_k).$$

Here are some examples of $g$.

**Example.** $g(x, y) = (x - y)^2$.

**Example.** $g(x, y) = |x - y|$.

**Example** (Wilcoxm's ranksom test). $g(x, y) = \mathbb{1}_{x_1 + x_2 > 0}$.

We're interested to know about $\mathbb{E}[g(X_1, X_2)]$. Assume $g$ is bounded by $B$, then

$$U(X) - U(X^{[k]}) \leq \frac{1}{\binom{n}{2}}(n-1)2B \leq \frac{4B}{n},$$

implying

$$\mathbb{P}(U - \mathbb{E}[U] \geq t) \leq e^{-\frac{nt^2}{8b^2}}$$

from McDiarmid's inequality with $c_i \coloneqq 2B$.

### Beyond McDiarmid's Inequality

Let's see some more advanced inequalities. In many cases, we want the variance to be small. While

$$\operatorname{Var}[X_1 + \cdots + X_n] \leq \sum_{i=1}^{n} \operatorname{Var}[X_i],$$

to have an inequality for a non-linear function, we have the following.

**Theorem 2.5.2** (Efron-Stein inequality)**.** Let $X_1, \ldots, X_n$ be independent random variables, and $X_1', \ldots, X_n'$ be i.i.d. copies of $X_i$'s. Then

$$\operatorname{Var}[f(X)] \leq \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}\left[\left(f(X) - f(X^{[i]})\right)^2\right].$$

**Proof.** See Theorem A.1.3 for a proof. ■

**Note.** We see that since $\operatorname{Var}[X] = \frac{1}{2}\mathbb{E}\left[(X - X')^2\right]$, by letting $f(X_1, \ldots, X_n) = \sum_i X_i$, if $f$ satisfies bounded condition, then $\operatorname{Var}[f] \leq \frac{1}{2}\sum_i c_i^2$.

Now, recall that by using McDiarmid's inequality, we can show that for $\mathscr{F} \ni f$ being $B$-bounded,

$$S_n \leq \mathbb{E}[S_n] + B\sqrt{\frac{2}{n}\log\frac{1}{\delta}}$$

with probability at least $1 - \delta$. However, what if the variance $\operatorname{Var}[f(X)]$ is small, but the maximum spread ($B$) is very large? In this case, we would want to replace $B$ in the inequality by $\operatorname{Var}[f(X)]$.

**Notation** (Empirical process notation)**.** Let $\mathbb{P}f = \mathbb{E}[f]$ and $\mathbb{P}_n f = \sum_i f(X_i)/n$.

This is achieved by the following, although it's much harder to prove [BLM13, §12].

**Theorem 2.5.3** (Talagrand's concentration inequality)**.** Let $\mathscr{F}$ is $B$-bounded, and $S_n = \sup_{f\in\mathscr{F}} |\mathbb{P}_n f - \mathbb{P}f|$. Then

$$S_n \leq c \cdot \mathbb{E}[S_n] + c\sqrt{\frac{\sup_{f\in\mathscr{F}} \operatorname{Var}[f(X_1)]}{n}\log\frac{1}{\alpha}} + c \cdot \frac{B}{n}\log\frac{1}{\alpha}$$

with probability at least $1 - \alpha$.

**Remark.** We might encounter an explicit situation where Talagrand's concentration is more profitable to use than bounded differences inequality later in the course.

# Chapter 3

# Expected Supremum of Empirical Process

## Lecture 6: A Glance at Statistical Learning Theory

6 Sep. 9:00

In this chapter, we're going to focus on solving Problem 1.3.1, i.e., bounding the supremum of the empirical process. Specifically, we're going to see two important techniques, one is chaining (which leads to the main Dudley's entropy bound), and another is bracketing (which leads to the bracketing bound). In both cases, we successfully bound the expected supremum of the empirical process.

## 3.1 Statistical Learning

### 3.1.1 Goodness of Fit Testing

Let's first see another motivation for studying the uniform law of large numbers, i.e., the *goodness of fit testing*. Given $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, we want to distinguish between $H_0 \colon \mathbb{P} = \mathbb{P}_0$ and $H_1 \colon \mathbb{P} \neq \mathbb{P}_0$.

Many tests are possible. One approach could be the Kolmogorov-Smirnov test: assume $F$ is the CDF of $\mathbb{P}_0$, then consider the Kolmogorov-Smirnov statistics:

**Definition 3.1.1** (Kolmogorov-Smirnov statistics)**.** The *Kolmogorov-Smirnov statistics* for a distribution $\mathbb{P}$ is defined as

$$D_n = \sup_{t \in \mathbb{R}} |F_n(t) - F(t)|$$

where $F_n(t)$ and $F$ is the empirical CDF and the CDF of $\mathbb{P}$, respectively.

From Glivenko-Cantelli theorem, $D_n \to 0$ under $H_0$, and $D_n$ should not converge to 0, under some alternative. Assuming continuity of $F$, Kolmogorov showed that

(a) the distribution $D_n$ does not depend on $F$;

(b) $D_n = O_p(1/\sqrt{n})$;

(c) $\sqrt{n} D_n \to \sup_{t \in [0,1]} |B(t)|$ where $B(t)$ is the Broweian bridge on $[0, 1]$.

(d) $\mathbb{P}(\sqrt{n} D_n \leq 2.4) \approx 0.999973$.

We'll take a non-asymptotic approach to this problem, i.e., we may not get such sharp constants.

### 3.1.2 Empirical Risk Minimization

Consider the following problem.

**Problem 3.1.1** (Empirical risk minimization)**.** Let $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$ be $n$ i.i.d. copies of $(X, Y) \in \mathcal{X} \times \mathscr{Y} \subseteq \mathbb{R}^d \times \mathbb{R}$ with distribution $\mathbb{P} = \mathbb{P}_X \times \mathbb{P}_{Y|X}$. Given a loss function $\ell \colon \mathscr{Y} \times \mathscr{Y} \to \mathbb{R}$

and a function class $\mathscr{F} = \{f : \chi \to \mathscr{Y}\}$, the *empirical risk minimization* is

$$\hat{f} \in \arg\min_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i).$$

**Example.** $\mathscr{F}$ can be the set of neural networks, decision trees, and linear functions.

**Example** (Linear regression)**.** Consider $\chi = \mathbb{R}^d$ and $\mathscr{Y} = \mathbb{R}$, with $\mathscr{F} = \{x \to w^\top x : w \in \mathbb{R}^d\}$ and $\ell(a, b) = (a - b)^2$.

**Example** (Linear classification)**.** Consider $\chi = \mathbb{R}^d$ and $\mathscr{Y} = \{0, 1\}$, with $\mathscr{F} = \{x \to (\text{sgn}(w^\top x) + 1)/2 : w \in B_2^d\}$ where $B_2^d$ is the unit ball in $d$-dimension, and $\ell(a, b) = \mathbb{1}_{a \neq b}$.

We also define the following.

**Definition.** Consider the set-up of empirical risk minimization.

**Definition 3.1.2** (Expected loss)**.** The *expected loss*[a] of $f \in \mathscr{F}$ is defined as

$$L(f) = \mathbb{E}_{(X,Y)\sim\mathbb{P}} \left[\ell(f(X), Y)\right].$$

[a] Also called *population loss* and *test error*.

**Definition 3.1.3** (Empirical loss)**.** The *empirical loss* is defined as

$$\hat{L}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i).$$

The main question in statistical learning is, what is an upper-bound on the expected loss of the empirical risk minimizer? If we plug in $\hat{f}$ instead of $f$, this is asking the test error of $\hat{f}$. To be more specific, $\hat{f}$ is basically a function of training data $S$, but when we look at

$$L(\hat{f}) = \mathbb{E}_{(X,Y)} \left[\ell(\hat{f}(X), Y)\right],$$

it is the expectation of future data points, i.e., it becomes a random variable, which is a function of $S$.

**Lemma 3.1.1.** For any $\mathscr{F}$, the empirical risk minimizer $\hat{f}$ satisfies

$$\mathbb{E}[L(\hat{f})] - \inf_{f \in \mathscr{F}} L(f) \leq \mathbb{E}\left[\sup_{f \in \mathscr{F}} \left(L(f) - \hat{L}(f)\right)\right].$$

**Proof.** Let $f^* = \arg\inf_{f \in \mathscr{F}} L(f)$. Then

$$L(\hat{f}) - L(f^*) = [L(\hat{f}) - \hat{L}(\hat{f})] + [\hat{L}(\hat{f}) - \hat{L}(f^*)] + [\hat{L}(f^*) - L(f^*)].$$

We see that

- $\hat{L}(\hat{f}) - \hat{L}(f^*) \leq 0$ by definition;
- $\hat{L}(f^*) - L(f^*) = 0$ in expectation since $f^*$ is fixed,
- We can't say $\mathbb{E}[L(\hat{f}) - \hat{L}(\hat{f})] = 0$ since $\hat{f}$ is also random.

Combine all these, we have

$$\mathbb{E}[L(\hat{f})] - \inf_{f\in\mathscr{F}} L(f) = \mathbb{E}[L(\hat{f}) - L(f^*)] \le \mathbb{E}[L(\hat{f}) - \hat{L}(\hat{f})] \le \mathbb{E}\left[\sup_{f\in\mathscr{F}} \left(L(f) - \hat{L}(f)\right)\right].$$

■

**Note.** Let us decode what Lemma 3.1.1 is claiming.

- Since $L(f)$ is the population error of $f$ and $\hat{L}(f)$ is the empirical loss of $f$, $\sup_{f\in\mathscr{F}} \left(L(f) - \hat{L}(f)\right)$ is the supremum of an empirical process.
- For the left-hand side, it represents the expected loss of $\hat{f}$ and the best possible out-of-sample error.[a] This is often called the excess risk.

[a] Or the best possible prediction error of $\mathscr{F}$.

**Definition 3.1.4** (Excess risk). The *excess risk* of an empirical risk minimizer $\hat{f}$ is given by

$$\mathbb{E}[L(\hat{f})] - \inf_{f\in\mathscr{F}} L(f).$$

**Remark.** For "curved" loss functions like square loss, supremum can be further "localized".

**Remark.** The bound in Lemma 3.1.1 can be vacuumed for now, e.g., for linear regression.

**Example** (1-D classification with thresholds). Let $\ell(a,b) = \mathbb{1}_{a\neq b} = a + (1-2a)b$ for $a, b \in \{0,1\}$. Then consider $a = y$ and $b = f(x)$,

$$\mathbb{E}\left[\sup_{f\in\mathscr{F}} \left(L(f) - \hat{L}(f)\right)\right] = \mathbb{E}\left[\sup_{f\in\mathscr{F}} \left(\mathbb{E}\left[Y + (1-2Y)f(X)\right] - \frac{1}{n}\sum_{i=1}^{n} \left(y_i + (1-2y_i)f(x_i)\right)\right)\right],$$

which can be viewed essentially as[a] the empirical process on the function $f$ instead of $\ell$,

$$\mathbb{E}\left[\sup_{f\in\mathscr{F}} \left(\mathbb{E}\left[f(X)\right] - \frac{1}{n}\sum_{i=1}^{n} f(x_i)\right)\right].$$

For 1-D case, assume that $\mathscr{F} = \{x \mapsto \mathbb{1}_{x\le\theta} : \theta \in \mathbb{R}\}$, then

$$\mathbb{E}\left[\sup_{\theta\in\mathbb{R}} \left(\mathbb{P}(X \le \theta) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x_i\le\theta}\right)\right] = \mathbb{E}\left[\sup_{\theta\in\mathbb{R}} (F(\theta) - F_n(\theta))\right],$$

i.e., $P(X \le \theta)$ is the CDF of the marginal distribution of $X$, $F(\theta)$, and $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x_i\le\theta}$ is the empirical CDF $F_n(\theta)$. Therefore, we go back to the same problem we introduced in the beginning of the chapter, i.e., the Kolmogorov-Smirnov statistics.

Let the term $\mathbb{P}(X \le \theta) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x_i\le\theta}$ to be a random variable $U_\theta$. One problem here is that we have infinitely many random variables, and they are also correlated with each other quite a lot. So how does this supremum behave?

Since each $U_\theta$ is at most 1, for any $\theta$, i.e., $\sup U_\theta \le 1$. So the worst case here is 1, and probably the best case is $O(1/\sqrt{n})$.

[a] Since $Y - \sum_i y_i/n$ is independent of $f$, so let's drop it; and $1-2Y$ is the sign, so can be dropped essentially.

# Lecture 7: Bracketing and Symmetrization

8 Sep. 9:00

Our main empirical process is so far $\mathbb{E}\left[\sup_{f\in\mathscr{F}} \mathbb{P}_n f - \mathbb{P}f\right]$. Let's first focus on the 1-D thresholds classification, i.e., we want to bound the supremum

$$\mathbb{E}\left[\sup_{\theta\in\mathbb{R}}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{x_i\le\theta} - \mathbb{P}(X\le\theta)\right|\right].$$

There are 2 approaches to bound this supremum: bracketing and symmetrization.

## 3.1.3 Bracketing

The main idea of bracketing is the following.

**Intuition.** Reduce an infinite number of random variables to finite, which will be more manageable.

Assume that $\mathbb{P}$ is continuous, and consider a finite set $\{\theta_i\}_{i=0}^{N+1}$ with $\theta_0 = -\infty$, $\theta_{N+1} = \infty$, such that they correspond to quantile of $\mathbb{P}$, i.e.,

$$\mathbb{P}(\theta_i \le X \le \theta_{i+1}) = \frac{1}{N+1}.$$

Given a $\theta$, $X$ will lie in between two adjacent $\theta_i$'s in the sequence. Denote the upper-bound as $u(\theta)$ and the lower-bound as $\ell(\theta)$ for this $\theta$, then

$$\begin{aligned}
\mathbb{P}(X\le\theta) - \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{x_i\le\theta} &\le \mathbb{P}(X\le u(\theta)) - \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{x_i\le\ell(\theta)}\\
&\le \mathbb{E}\left[\mathbb{1}_{X\le u(\theta)}\right] - \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{x_i\le\ell(\theta)}\\
&\le \mathbb{E}\left[\mathbb{1}_{X\le\ell(\theta)}\right] - \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{x_i\le\ell(\theta)} + \mathbb{P}(\ell(\theta)\le X\le u(\theta))\\
&\le \mathbb{E}\left[\mathbb{1}_{X\le\ell(\theta)}\right] - \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{x_i\le\ell(\theta)} + \frac{1}{N+1}
\end{aligned}$$

if we take the supremum over $\ell(\theta)\in\mathbb{R}$ instead of $\theta$,

$$\le \frac{1}{N+1} + \mathbb{E}\left[\max_{0\le j\le N}\mathbb{E}\left[\mathbb{1}_{X\le\theta_j}\right] - \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{x_i\le\theta_j}\right]. \tag{3.1}$$

To further bound Equation 3.1, recall the following.

**As previously seen.** If $X_i\sim\text{Subg}(\sigma^2)$ independent, $\sum_i a_iX_i\sim\text{Subg}((\sum_i a_i^2)\sigma^2)$ from Lemma 2.3.3.

**Remark.** Let $a_i = 1/n$, we see that $\mathbb{E}\left[\mathbb{1}_{X\le\theta_j}\right] - \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{x_i\le\theta_j}\in\text{Subg}(1/n)$.[a]

[a] Since it's bounded between 0 and 1.

Then, we can show the final bound.

**Proposition 3.1.1** (Bracketing). Let $x_1,\dots,x_n\overset{\text{i.i.d.}}{\sim}\mathbb{P}$, and $\mathscr{F} = \{\mathbb{1}_{X\le\theta}\colon\theta\in\mathbb{R}\}$. Then

$$\mathbb{E}_X\left[\sup_{f\in\mathscr{F}}\left(\mathbb{P}(X\le\theta) - \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{x_i\le\theta}\right)\right] = O\left(\sqrt{\frac{\log n}{n}}\right).$$

**Proof.** From Lemma 2.3.4, since we have $(N+1)$ random variables with variance factor $1/n$, by

choosing $N+1 \coloneqq n$,[a] Equation 3.1 can be further bounded by

$$\sqrt{\frac{2\log(N+1)}{n}} + \frac{1}{N+1} = O\left(\sqrt{\frac{\log n}{n}}\right).$$

■

[a] Recall that $n$ is the sample size, so we can choose the corresponding $n$ to meet the requirement.

### 3.1.4 Symmetrization

Another technique is called symmetrization which is essentially stated in the following lemma.

**Lemma 3.1.2** (Symmetrization). Given a function class $\mathscr{F} = \{f : \chi \to \mathscr{Y}\}$ and $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, and $\epsilon_1, \dots, \epsilon_n$ be i.i.d. Rademacher random variables. Then

$$\max\left(\mathbb{E}\left[\sup_{f\in\mathscr{F}} \mathbb{P}_n f - \mathbb{P}f\right], \mathbb{E}\left[\sup_{f\in\mathscr{F}} \mathbb{P}f - \mathbb{P}_n f\right]\right) \le 2\mathbb{E}_{\epsilon_i, X_i}\left[\sup_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right].$$

In particular,

$$\mathbb{E}\left[\sup_{f\in\mathscr{F}} |\mathbb{P}_n f - \mathbb{P}f|\right] \le 2\mathbb{E}_{\epsilon_i, X_i}\left[\sup_{f\in\mathscr{F}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right|\right].$$

**Proof.** Let $X_i'$'s be i.i.d. copies of $X_i$'s for all $i$. Since adding a sign $\epsilon_i$ won't change the expectation,[a]

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathscr{F}} \mathbb{E}\left[f(X)\right] - \frac{1}{n}\sum_{i=1}^n f(X_i)\right] &= \mathbb{E}\left[\sup_{f\in\mathscr{F}} \mathbb{E}_{X_i'}\left[\frac{1}{n}\sum_{i=1}^n f(X_i') - \frac{1}{n}\sum_{i=1}^n f(X_i)\right]\right] \\
&\le \mathbb{E}_{X_i}\left[\mathbb{E}_{X_i'}\left[\sup_{f\in\mathbb{F}} \frac{1}{n}\sum_{i=1}^n (f(X_i') - f(X_i))\right]\right] \\
&= \mathbb{E}_{X_i, X_i', \epsilon_i}\left[\sup_{f\in\mathbb{F}} \frac{1}{n}\sum_{i=1}^n (f(X_i') - f(X_i))\epsilon_i\right] \\
&\le \mathbb{E}_{X_i', \epsilon_i}\left[\sup_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^n f(X_i')\epsilon_i\right] + \mathbb{E}_{X_i, \epsilon_i}\left[\sup_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^n f(X_i)\epsilon_i\right] \\
&= 2\mathbb{E}\left[\sup_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right].
\end{aligned}$$

■

[a] Since the distributions of $f(X_i') - \sum_i f(X_i)$ and $f(X_i) - \sum_i f(X_i')$ are the same.

**Intuition.** If we condition on $X_i$'s, the bound can be seen as a linear combination of Rademacher random variables. Thus, we can refer to properties of sub-Gaussian random variables.

The upper-bound deserves a special name.

**Definition 3.1.5** (Rademacher complexity). Let $X_i \overset{\text{i.i.d.}}{\sim} \mathbb{P}$ be independent and $\epsilon_i$ be i.i.d. Rademacher random variables. The *Rademacher complexity* of a function class $\mathscr{F}$ w.r.t. $\mathbb{P}$ is

$$R_n(\mathscr{F}) \coloneqq \mathbb{E}_{\epsilon_i, X_i}\left[\sup_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right].$$

On the other hand, the opposite direction of symmetrization lemma also holds.

**Lemma 3.1.3.** Given a function class $\mathscr{F} = \{f : \chi \to \mathscr{Y}\}$ and $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, and $\epsilon_1, \dots, \epsilon_n$ be i.i.d. Rademacher random variables. Then

$$\mathbb{E}_{X_i,\epsilon_i}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right|\right] \le 2\mathbb{E}\left[\sup_{f\in\mathscr{F}}|\mathbb{P}_n f - \mathbb{P}f|\right] + \frac{1}{\sqrt{n}}\sup_{f\in\mathscr{F}}|\mathbb{P}f|.$$

**Proof.** This technique is so-called *desymmetrization*: Consider

$$\begin{aligned}
&\mathbb{E}_{\epsilon_i,X_i}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right|\right] \\
&\le \mathbb{E}_{\epsilon_i,X_i}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i (f(X_i) - \mathbb{E}\left[f(X)\right])\right|\right] + \mathbb{E}_{\epsilon_i}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \mathbb{E}\left[f(X)\right]\right|\right] \\
&= \mathbb{E}_{\epsilon_i,X_i,X_i'}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i (f(X_i) - \mathbb{E}\left[f(X_i')\right])\right|\right] + \mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \mathbb{E}_{\epsilon_i}\left[f(X_i)\right]\right|\right].
\end{aligned}$$

The first term can be further bounded by

$$\begin{aligned}
\mathbb{E}_{\epsilon_i,X_i,X_i'}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i (f(X_i) - \mathbb{E}\left[f(X_i')\right])\right|\right] &\le \mathbb{E}_{\epsilon_i,X_i,X_i'}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i (f(X_i) - f(X_i'))\right|\right] \\
&= \mathbb{E}\left[\sup_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^\infty (f(X_i) - f(X_i'))\right] \\
&= \mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \left(f(X_i) - f(X_i') + (\mathbb{E}\left[f\right] - \mathbb{E}\left[f\right])\right)\right|\right] \\
&= 2\mathbb{E}\left[\sup_{f\in\mathscr{F}}|\mathbb{P}_n f - \mathbb{P}f|\right],
\end{aligned}$$

and the second term can be bounded by

$$\mathbb{E}_{\epsilon_i}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \mathbb{E}\left[f(X)\right]\right|\right] \le \sup_{f\in\mathscr{F}}|\mathbb{E}\left[f(X)\right]| \cdot \mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\right|\right] \le \frac{1}{\sqrt{n}}\sup_{f\in\mathscr{F}}|\mathbb{P}f|$$

where $\mathbb{E}\left[\left|\frac{1}{n}\sum_i \epsilon_i\right|\right] \le \frac{c}{\sqrt{n}}$ with $c = 1$. Combine them together, we have the final result. ■

# Lecture 8: Symmetrization on 1-D Threshold Classification

Analogous to the Rademacher complexity defined for a function class w.r.t. $\mathbb{P}$, we can define it on a set. 11 Sep. 9:00

**Definition 3.1.6** (Rademacher width)**.** Let $\epsilon_i$ be i.i.d. Rademacher random variables. Then the *Rademacher width*[a] of a set $A \subseteq \mathbb{R}^n$ is defined as

$$R_n(A) = \mathbb{E}_{\epsilon_i}\left[\sup_{a\in A}\frac{1}{n}\sum_{i=1}^n \epsilon_i a_i\right].$$

[a] Also called *Rademacher average*.

**Notation.** People sometimes just say "Rademacher complexity" for Rademacher width.

Now, applying the symmetrization lemma to $\mathscr{F} = \{\mathbb{1}_{X\le\theta} : \theta \in \mathbb{R}\}$, we have the following result that is comparable to Proposition 3.1.1.

**Proposition 3.1.2.** Let $x_1, \dots, x_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, and $\mathscr{F} = \{\mathbb{1}_{x \le \theta} \colon \theta \in \mathbb{R}\}$. Then

$$\mathbb{E}_X \left[ \sup_{f \in \mathscr{F}} \left( \mathbb{P}(X \le \theta) - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{x_i \le \theta} \right) \right] = O\left( \sqrt{\frac{\log n}{n}} \right).$$

**Proof.** From the symmetrization lemma,

$$\begin{aligned} \mathbb{E}_{X, x_i} \left[ \sup_{\theta \in \mathbb{R}} \left( \mathbb{P}(X \le \theta) - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{x_i \le \theta} \right) \right] &\le 2 \mathbb{E}_{\epsilon_i, x_i} \left[ \sup_{\theta \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n \epsilon_i \mathbb{1}_{x_i \le \theta} \right] \quad \text{condition on } x_1, \dots, x_n \\ &= 2 \mathbb{E}_{x_i} \left[ \mathbb{E}_{\epsilon_i | x_i} \left[ \sup_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_i \mathbb{1}_{x_i \le \theta} \middle| x_1, \dots, x_n \right] \right]. \end{aligned}$$

Let $V_\theta \coloneqq \frac{1}{n} \sum_{i=1}^n \epsilon_i \mathbb{1}_{x_i \le \theta}$, we see that there are only $n+1$ distinct $V_\theta$'s, and it's constant in the intervals $\theta \in [X_{(k)}, X_{(k+1)})$ for $k = 0, \dots, n-1$ where $X_{(k)}$ are the order statistics with $X_{(0)} \coloneqq -\infty$. Now, define $\theta_k \coloneqq X_{(k)}$, we can then write

$$\sup_{\theta \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n \epsilon_i \mathbb{1}_{x_i \le \theta} = \max_{k = 0, \dots, n} \frac{1}{n} \sum_{i=1}^n \epsilon_i \mathbb{1}_{x_i \le \theta_k},$$

hence,

$$2 \mathbb{E}_{x_i} \left[ \mathbb{E}_{\epsilon_i | x_i} \left[ \sup_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_i \mathbb{1}_{x_i \le \theta} \middle| x_1, \dots, x_n \right] \right] = 2 \mathbb{E}_{x_i} \left[ \mathbb{E}_{\epsilon_i | x_i} \left[ \max_{k=0,\dots,n} V_{\theta_k} \middle| x_1, \dots, x_n \right] \right]$$

with $V_{\theta_k} \sim \text{Subg}(1/n)$ and Lemma 2.3.4,

$$\le 2 \mathbb{E}_{x_i} \left[ \sqrt{\frac{2}{n} \log(n+1)} \right] = O\left( \sqrt{\frac{\log n}{n}} \right).$$

■

**Remark.** Looking back to the example of 1-D thresholds classification, we see that the excess risk of ERM is $O(\sqrt{\log n / n})$.

## 3.2 Vapnik-Chervonenkis Dimension

### 3.2.1 Glivenko-Cantelli Class

From bracketing and symmetrization, we see that there are classes of functions such that

$$\sup_{f \in \mathscr{F}} |\mathbb{P} f - \mathbb{P}_n f| \to 0$$

as $n \to \infty$. They deserve their own name.

**Definition 3.2.1** (Glivenko-Cantelli). A function class $\mathscr{F} = \{f \colon \chi \to \mathbb{R}\}$ is *Glivenko-Cantelli* w.r.t. $\mathbb{P}$ if as $n \to \infty$,

$$\sup_{f \in \mathscr{F}} |\mathbb{P} f - \mathbb{P}_n f| \to 0.$$

From bracketing and symmetrization, we know that $\mathscr{F} = \{\mathbb{1}_{X \le \theta} \colon \theta \in \mathbb{R}\}$ is Glivenko-Cantelli. Let's see some counterexamples.

**Example.** Let $\chi = \mathbb{R}$, $\mathscr{F} = \{\mathbb{1}_A \colon A \subseteq \chi, |A| < \infty\}$, and $\mathbb{P}$ be any continuous measure on $\chi$. Then $\mathscr{F}$ is not Glivenko-Cantelli w.r.t. $\mathbb{P}$.

**Proof.** For $f = \mathbb{1}_A$, $\mathbb{P}f = \mathbb{P}(X \in A) = 0$ since $|A| < \infty$. On the other hand, let $A_0 = \{X_1, \dots, X_n\}$ be the observed empirical data, $\mathbb{P}_n f = 1$, i.e., $\sup_{f \in \mathscr{F}} |\mathbb{P}f - \mathbb{P}_n f| = 1$ for all $n \in \mathbb{N}$. ⊛

**Example.** Let $\chi = \mathbb{R}$, $\mathscr{F} = \{f : \chi \to \mathbb{R} \text{ bounded and continuous}\}$, and $\mathbb{P} = \mathcal{U}[0,1]$. Then $\mathscr{F}$ is not Glivenko-Cantelli.

**Proof.** Consider $f(X_i) = 1$ for $i = 1, \dots, n$ and $f = 0$ elsewhere (continuously),[a] then we can make $\int_0^1 f(t)\,\mathrm{d}t < \delta$ for some $\delta \in (0,1)$. This implies $\sup_{f \in \mathscr{F}} |\mathbb{P}f - \mathbb{P}_n f| \geq 1 - \delta$ for all $n \in \mathbb{N}$. ⊛

[a] E.g., sharp peak at $X_i$'s.

### 3.2.2 Vapnik-Chervonenkis Dimension

**Notation.** Let $\mathscr{F}(x_1, \dots, x_n) \coloneqq \{(f(x_1), \dots, f(x_n))\}_{f \in \mathscr{F}} \subseteq \mathbb{R}^n$.

We can relate the Rademacher width of $\mathscr{F}(X_1, \dots, X_n)$ to the Rademacher complexity of $\mathscr{F}$ since[1]

$$\mathbb{E}_{X_i}\left[R_n(\mathscr{F}(X_1, \dots, X_n))\right] = \mathbb{E}_{X_i, \epsilon_i}\left[\sup_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i)\right] = R_n(\mathscr{F}).$$

Moreover, we see that if $\mathscr{F}(X_1, \dots, X_n)$ is finite, by the same proof as in Proposition 3.1.2,

$$\mathbb{E}_{X_i}\left[R_n(\mathscr{F}(X_1, \dots, X_n))\right] \leq 2\sqrt{\frac{2\log|\mathscr{F}(X_1, \dots, X_n)|}{n}}.$$

The up-shot is the following.

**Remark.** If $|\mathscr{F}(X_1, \dots, X_n)| \leq n^d$ for some $d \in \mathbb{N}^+$, then we again get an $O(\sqrt{\log n / n})$ bound.

This is captured by the polynomial discrimination, where we're going to focus on boolean functions.

**Definition 3.2.2** (Polynomial discrimination)**.** We say that a boolean function class $\mathscr{F}$ on $\chi$ has a *polynomial discrimination* if for all $x_1, \dots, x_n \in \chi$, $|\mathscr{F}(x_1, \dots, x_n)| \leq \mathsf{poly}(n)$.

To characterize $|\mathscr{F}(x_1, \dots, x_n)|$, we will look at the VC dimension of $\mathscr{F}$, which is related to the size of the discrimination of $\mathscr{F}$ in a non-trivial way.

**Definition.** Let $\mathscr{F}$ be a boolean function class on $\chi$.

**Definition 3.2.3** (Shatter)**.** A finite set $\{x_1, \dots, x_D\} \subseteq \chi$ is *shattered* by $\mathscr{F}$ if $\mathscr{F}(x_1, \dots, x_D) = \{0,1\}^D$.[a]

[a] We take the convention that $\varnothing$ is always shattered.

**Definition 3.2.4** (Vapnik-Chervonenkis dimension)**.** The *VC dimension* of $\mathscr{F}$ on $\chi$ is the maximum integer $D$ such that there exists a size $D$ finite set $A \subseteq \chi$ shattered by $\mathscr{F}$.

Let's consider some examples on $\chi = \mathbb{R}$.

**Example.** The VC dimension of $\mathscr{F} = \{\mathbb{1}_{X \leq \theta} : \theta \in \mathbb{R}\}$ is 1.

**Example.** The VC dimension of $\mathscr{F} = \{\mathbb{1}_{[a,b]} : a, b \in \mathbb{R}\}$ is 2.

Let's look at one example with $\chi = \mathbb{R}^2$.

[1] This is why people overload $R_n$ for both Rademacher width and Rademacher complexity.

**Example.** The VC dimension of $\mathscr{F} = \{\mathbb{1}_{[a,b]\times[c,d]} : a, b, c, d \in \mathbb{R}\}$ is 4.

# Lecture 9: VC Dimension

Given VC dimension, we can upper-bound the size of the discrimination. We need Pajor's lemma first. 13 Sep. 9:00

**Lemma 3.2.1** (Pajor's lemma). Given a boolean function class $\mathscr{F}$ on a finite set $\Omega$, then

$$|\mathscr{F}| \le |\{S \subseteq \Omega : S \text{ shattered by } \mathscr{F}\}|.$$

**Proof.** We prove this by induction on $n$. For $n = 1$ (base case), it holds trivially since

$$|\mathscr{F}| = 2 \le |\{S \subseteq \Omega : S \text{ shattered by } \mathscr{F}\}|.$$

Assume the statement holds for all $\Omega$ such that $|\Omega| = n$. For $|\Omega| = n+1$, we write $\Omega = (\Omega \setminus \{x_0\}) \cup \{x_0\} =: \Omega_0 \cup \{x_0\}$ and let $\mathscr{F}_0$ and $\mathscr{F}_1$ be two boolean function classes defined on $\Omega_0$ as

$$\mathscr{F}_0 = \{f \in \mathscr{F} : f(x_0) = 0\}, \quad \mathscr{F}_1 = \{f \in \mathscr{F} : f(x_0) = 1\}.$$

We further define $S_{\mathscr{F}'}$ as $S_{\mathscr{F}'} = \{S \subseteq \Omega' : S \text{ shattered by } \mathscr{F}'\}$ for any function class $\mathscr{F}'$ defined on $\Omega'$. Then, by induction hypothesis, $|\mathscr{F}_i| \le |S_{\mathscr{F}_i}|$, hence

$$|\mathscr{F}| = |\mathscr{F}_0| + |\mathscr{F}_1| \le |S_{\mathscr{F}_0}| + |S_{\mathscr{F}_1}|.$$

Finally, we claim the following.

**Claim.** $|S_{\mathscr{F}_0}| + |S_{\mathscr{F}_1}| \le |S_{\mathscr{F}}|$.

**Proof.** Let $S \subseteq \Omega_0$ shattered by both $\mathscr{F}_0$ and $\mathscr{F}_1$, then $S$ is shattered by $\mathscr{F}$ too. Moreover, Observe that $S \cup \{x_0\}$ is shattered by $\mathscr{F}$ but not $\mathscr{F}_i$ ($f(x_0)$ is fixed for $f \in \mathscr{F}_i$). Now, when

- $S$ is shattered by only one of the $\mathscr{F}_i$'s: $S$ contributes one unit both to $|S_{\mathscr{F}}|$ and $|S_{\mathscr{F}_i}|$;
- $S$ is shattered by both $\mathscr{F}_i$'s, $S$ and $S \cup \{x_0\}$ are shattered by $\mathscr{F}$: $S$ contributes two units to $|S_{\mathscr{F}}|$ and one unit to both $|S_{\mathscr{F}_i}|$'s.

By counting, we're done (it's possible that $S$ is shattered by $\mathscr{F}$ but not $\mathscr{F}_i$'s, so $\le$). ⊛

This implies $|\mathscr{F}| \le |S_{\mathscr{F}}|$ for $|\Omega| = n+1$, i.e., the induction is done. ■

We can then prove the Sauer-Shelah lemma.

**Lemma 3.2.2** (Sauer-Shelah lemma). Let $\mathscr{F}$ be a boolean function class such that $\mathrm{VC}(\mathscr{F}) = d$, then for every $\{x_1, \dots, x_n\} \subseteq \chi$ such that $n \ge d$,

$$|\mathscr{F}(x_1, \dots, x_n)| \le \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{d} \le \left(\frac{en}{d}\right)^d.$$

**Proof.** Let $\Omega$ be a set of size $n$, then the number of subsets with size $\le d$ is $\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{d}$, hence by the definition of VC dimension,

$$|\{S \subseteq \Omega : S \text{ shattered by } \mathscr{F}\}| \le \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{d}.$$

With the standard Stirling's estimation gives the result. ■

Then, as our motivation suggests, the same proof of Proposition 3.1.2 applies, giving the bound on the Rademacher complexity in terms of the VC dimension.

**Proposition 3.2.1.** For any boolean function class $\mathscr{F}$, if $n \geq \mathrm{VC}(\mathscr{F})$, for some constant $c > 0$,

$$R_n(\mathscr{F}) \leq c\sqrt{\frac{\mathrm{VC}(\mathscr{F})}{n}\log\left(\frac{en}{\mathrm{VC}(\mathscr{F})}\right)}.$$

**Remark.** We see that Proposition 3.2.1 is independent of $\mathbb{P}$, i.e., the bounds still holds after taking $\sup_{\mathbb{P}}$ on the left-hand side. However, if $\mathrm{VC}(\mathscr{F}) = \infty$, then this "distribution-free" uniform convergence fails.

However, if we don't care about the distribution-free property, we do have examples that the uniform convergence holds for a particular $\mathbb{P}$ when $\mathrm{VC}(\mathscr{F}) = \infty$.

**Example.** For $\mathscr{F} = \{\mathbb{1}_A \colon \text{compact convex } A \subseteq [0,1]^d\}$, $\mathrm{VC}(\mathscr{F}) = \infty$. If $\mathbb{P}$ is continuous w.r.t. Lebesgue's measure, then the uniform law of large numbers still holds.

**Remark.** The $\sqrt{\log n}$ factors in Proposition 3.2.1 is superfluous (Corollary 3.3.5).

**Example.** Let $V$ be a vector space of real function on $\chi$ with $\dim(V) = D$, and $\mathscr{F} = \{\mathbb{1}_{f\geq 0} \colon f \in V\}$. Then $\mathrm{VC}(\mathscr{F}) \leq D$.

**Proof.** We want to show that for any $\{x_1, \ldots, x_{D+1}\}$ can't be shatterred. Let

$$T = \{(f(x_1), \ldots, f(x_{D+1})) \colon f \in V\},$$

which is a linear subspace of $\mathbb{R}^{D+1}$ such that $\dim(T) \leq D$. Hence, there exists a non-zero $y \in \mathbb{R}^{D+1}$ such that $\sum_{i=1}^{D+1} y_i f(x_i) = 0$ for all $f \in V$. Now, without loss of generality, there exists an index $k$ such that $y_k > 0$. If $\mathscr{F}$ shatters $\{x_1, \ldots, x_{D+1}\}$, then there exists $f \in V$ such that

$$\begin{cases} f(x_i) < 0, & \forall i \colon y_i > 0; \\ f(x_i) \geq 0, & \forall i \colon y_i \leq 0. \end{cases}$$

But then $\sum_i y_i f(x_i) < 0$, which is a contradiction. ⊛

**Example** (Half-space). For $\mathscr{F} = \{\mathbb{1}_H \colon \text{half space } H \subseteq \mathbb{R}^d\}$, $\mathrm{VC}(\mathscr{F}) = d + 1$.

Although it seems like $\mathrm{VC}(\mathscr{F}) \approx \#\text{parameters of } \mathscr{F}$; however, it's not true in general.

**Example.** Consider $\mathscr{F} = \{x \mapsto \mathbb{1}_{\sin tx \geq 0} \colon t \in \mathbb{R}^+\}$, then $\mathrm{VC}(\mathscr{F}) = \infty$.

# Lecture 10: Discretization of a Space

## 3.3 Metric Entropy Methods

15 Sep. 9:00

We have been focusing on the boolean function class with finite VC dimension, and our goal now is to generalize beyond the boolean case. This can be done by discretizing of a space.

**Intuition** (Informal principle). We want to bound $\mathbb{E}\left[\sup_{t\in T} X_t\right]$. If $\{X_t\}_{t\in T}$ is "sufficiently continuous", then $\mathbb{E}\left[\sup_{t\in T} X_t\right]$ is governed by metric properties of $T$ (metric entropy!).

First, we need the notion of pseudo-metric from the analysis.

**Definition 3.3.1** (Pseudo-metric). Given a space $T$, a function $d\colon T \times T \to \mathbb{R}^+$ is a *pseudo-metric* if

(a) $d(x,x) = 0$ for all $x \in T$;[a]

(b) $d(x,y) = d(y,x)$ for all $x, y \in T$;

(c) $d(x,y) \le d(x,z) + d(y,z)$ for all $x, y, z \in T$.

[a] If $d$ further satisfies that $d(x,y) > 0$ for all $x \neq y$, then it becomes a *metric*.

**Note.** The motivation for looking at pseudo-metric instead of the usual metric is because, considering observed data $x_1, \dots, x_n$ at hands, the most natural distance might be

$$(f,g) \mapsto \sqrt{\frac{1}{n}\sum_{i=1}^{n} \big(f(x_i) - g(x_i)\big)^2},$$

which is a pseudo-metric since $f$ and $g$ can agree only on $x_i$'s and vary elsewhere.

### 3.3.1 Covering Number and Packing Number

Now, let $(T,d)$ denote a pseudo-metric space in the remaining of this section, unless specified.

**Definition 3.3.2** ($\epsilon$-net). A set $N$ is an *$\epsilon$-net* of $(T,d)$ if for all $t \in T$, there exists $\pi(t) \in N$ such that $d(t,\pi(t)) \le \epsilon$.

**Definition 3.3.3** (Covering number). The *$\epsilon$-covering number* $N(T,d,\epsilon)$ of $(T,d)$ is defined as

$$N(T,d,\epsilon) \coloneqq \inf\{|N| \colon N \text{ is an } \epsilon\text{-net for } (T,d)\}.$$

**Remark.** $N$ is not necessary a subset of $T$ for convenience. Furthermore, if $N \nsubseteq T$, one can construct another net $N'$ such that $N' \subseteq T$ and $N'$ is a $2\epsilon$-net.

**Definition 3.3.4** (Totally bounded). $(T,d)$ is *totally bounded* if for all $\epsilon > 0$, $N(T,d,\epsilon) < \infty$.

**Definition 3.3.5** ($\epsilon$-packing). A set $N \subseteq T$ is an *$\epsilon$-packing* of $(T,d)$ if for all $t \neq t'$ in $N$, $d(t,t') > \epsilon$.

**Definition 3.3.6** (Packing number). The *$\epsilon$-pacing number* $M(T,d,\epsilon)$ of $(T,d)$ is defined as

$$M(T,d,\epsilon) = \sup\{|N| \colon N \text{ is an } \epsilon\text{-packing of } (T,d)\}.$$

As the title suggests, we define the following metric properties, which is an essential notion that helps us to bound the expected empirical process supremum.

**Definition 3.3.7** (Metric entropy). The *metric entropy* of $(T,d)$ is defined as $\log M(T,d,\epsilon)$.

The fact that we're using packing number $M(T,d,\epsilon)$ when defining metric entropy is not relevant here due to the following.

**Lemma 3.3.1.** Given a $(T,d)$, for any $\epsilon > 0$,

$$M(T,d,2\epsilon) \le N(T,d,\epsilon) \le M(T,d,\epsilon).$$

**Proof.** We show them one by one.

**Claim.** $M(T, d, 2\epsilon) \leq N(T, d, \epsilon)$.

**Proof.** Take $\mathcal{M}$ to be a $2\epsilon$-packing and $\mathcal{N}$ to be an $\epsilon$-net. Then for any $t \in \mathcal{N}$, consider $B(t, \epsilon)$. We see that there is at most one $x \in \mathcal{M}$ such that $d(t, x) \leq \epsilon$ since otherwise, if $x, x' \in \mathcal{M}$ such that $x \neq x'$ and $d(t, x), d(t, x') \leq \epsilon$, then $d(x, x') \leq 2\epsilon$, a contradiction to $\mathcal{M}$. ⊛

**Claim.** $N(T, d, \epsilon) \leq M(T, d, \epsilon)$.

**Proof.** Take $\mathcal{M}$ to be a maximum $\epsilon$-packing, it suffices to show that $\mathcal{M}$ is also an $\epsilon$-net, i.e., for all $t \in T$, there exists $x \in \mathcal{M}$ such that $d(x, t) \leq \epsilon$. Suppose not, then $d(t, x) > \epsilon$ for all $x \in \mathcal{M}$, i.e., we can add $x$ to $\mathcal{M}$, contradiction. ⊛

■

For simplicity, we will use the following notations.

**Notation.** If $(T, d)$ and $\epsilon$ are clear from the context, we write $N := N(T, d, \epsilon)$ and $M := M(T, d, \epsilon)$.

Turns out that there's a characterization of the packing number of the unit ball in Euclidean space.

**Proposition 3.3.1.** Consider $(\mathbb{R}^d, \|\cdot\|)$ where $\|\cdot\|$ is any norm. Denote $B = \{x \colon \|x\| \leq 1\}$, then for all $\epsilon > 0$,

$$(1/\epsilon)^d \leq M(B, \|\cdot\|, \epsilon) \leq (1 + 2/\epsilon)^d.$$

**Proof.** For the lower-bound, we see that

$$N \operatorname{vol}(\epsilon B) \geq \operatorname{vol}(B) \Rightarrow N\epsilon^d \geq 1.$$

With $N \leq M$ from Lemma 3.3.1, we get the lower-bound.

For the upper-bound, since $\epsilon/2$ balls around points in $M$ are disjoint, union of these $\epsilon/2$ balls will lie in $(1 + \epsilon/2)B$. This implies

$$M \times \left(\frac{\epsilon}{2}\right)^d \times \operatorname{vol}(B) \leq \left(1 + \frac{\epsilon}{2}\right)^d \times \operatorname{vol}(B) \Rightarrow M \leq \left(1 + \frac{2}{\epsilon}\right)^d.$$

■

**Note.** From Proposition 3.3.1, $\log M(\mathbb{R}^d, \|\cdot\|, \epsilon) \approx d \log 1/\epsilon$.

### 3.3.2 Hölder Smooth Functions

We are interested in looking at function spaces, and the following are the canonical smooth function classes studied in *nonparametric regression*.

**Definition 3.3.8** (Hölder smooth function class)**.** Let $\alpha > 0$ and $\beta = \lfloor \alpha \rfloor$. Then the *Hölder smooth function class* $\mathcal{S}_\alpha$ is defined to be the class of functions on $[0, 1]$ such that

(a) $f$ continuous on $[0, 1]$;

(b) $f$ is $\beta$-times differentiable;

(c) $|f^{(k)}| \leq 1$ for all $k = 0, \ldots, \beta$;

(d) $|f^{(\beta)}(x) - f^{(\beta)}(y)| \leq |x - y|^{\alpha - \beta}$ for all $x, y \in [0, 1]$.

**Note.** When $\alpha = 1$, $\mathcal{S}_\alpha$ is a class of 1-Lipschitz functions.

**Remark.** The Hölder smooth function classes are nested, so it's not surprising that the metric entropies decrease as $\alpha$ increases.

Now, let $d(f,g) = \sup_{x\in[0,1]} |f(x) - g(x)|$, then $(\mathcal{S}_\alpha, d)$ is a pseudo-metric space.

**Theorem 3.3.1.** There exists constants $c_1, c_2 > 0$ only depend on $\alpha$ such that for all $\epsilon > 0$,

$$\exp\left(c_2\epsilon^{-1/\alpha}\right) \le M(\mathcal{S}_\alpha, d, \epsilon) \le \exp\left(c_1\epsilon^{-1/\alpha}\right).$$

**Proof sketch.** Here we illustrate the basic idea when $\alpha = 1$, i.e., the set of $[0,1]$ valued 1-Lipschitz functions on $[0,1]$. We only sketch the proof of the upper-bound, since the lower-bound is similar.

Firstly, we partition both the domain and the range of $f$ with small intervals with width $\epsilon$, resulting in $1/\epsilon$ small intervals on both the $x$-axis and the $y$-axis.

Take any function $f \in \mathscr{F}$. We construct a piece-wise constant function $\widetilde{f}$ which approximates $f$. On each small interval in the $x$-axis, we can define $\widetilde{f}$ to be constant, taking value equal to the midpoint of the interval in the $y$-axis where the value of $f$ at the left endpoint of this interval (in the $x$-axis) lies. Then, we have the following.

**Claim.** $\sup_{x\in[0,1]} |f(x) - \widetilde{f}(x)| \le C\epsilon$.

**Proof.** Since $f$ cannot vary by more than $\epsilon$ in any interval of length $\epsilon$. ⊛

Now, as we vary $f \in \mathscr{F}$, consider the following.

**Problem.** What is the number of distinct $\widetilde{f}$ we can get?

A trivial bound is that, in each small interval on the $x$-axis, it takes one of the midpoints of the intervals on the $y$-axis, and hence, the number of such functions is bounded by $(\frac{1}{\epsilon})^{\frac{1}{\epsilon}}$.

We can do slightly better. Note that, for the first interval, the number of possible values of $\widetilde{f}$ is $\frac{1}{\epsilon}$. However, after that, in the next interval, the value of $\widetilde{f}$ can only go up one interval, down one interval, or stay the same (due to 1-Lipschitzness of $f$), i.e., there are only 3 choices afterward for every interval, going from left to right, resulting an upper-bound on the number of distinct $\widetilde{f}$ as

$$\frac{1}{\epsilon} 3^{\frac{1}{\epsilon}-1} \le \exp\left(\frac{C}{\epsilon}\right).$$

For a complete (and long) proof, see Theorem A.2.2. ■

**Remark.** Comparing Proposition 3.3.1 and Theorem 3.3.1, we see that the metric entropy is logarithmic in $1/\epsilon$ versus some exponent of $1/\epsilon$. This is typically the hallmark of a parametric versus a nonparametric function class.

# Lecture 11: Gaussian and Sub-Gaussian Process

## 3.3.3 Sub-Gaussian Process

18 Sep. 9:00

**As previously seen.** Given a stochastic process $\{X_t\}_{t\in T}$ with $(T,d)$, we want to bound $\mathbb{E}\left[\sup_{t\in T} X_t\right]$.

Recall our informal principle, i.e., if $\{X_t\}_{t\in T}$ is "sufficiently continuous" w.r.t. $d$, then $\mathbb{E}\left[\sup_{t\in T} X_t\right]$ is governed by metric properties (e.g., metric entropy) of $T$. We start by considering the Gaussian process.

**Definition 3.3.9** (Gaussian process). A stochastic process $\{X_t\}_{t\in T}$ is a *Gaussian process* if for any finite set of indices $t_1, \dots, t_k$, $(X_{t_1}, \dots, X_{t_k}) \sim \mathcal{N}(0, \mathbf{\Sigma})$.

Clearly, this is a very strong notion due to the following.

**Note.** For $d(t,t') = \sqrt{\mathbb{E}\left[(X_t - X_{t'})^2\right]}$, we have

$$\mathbb{E}\left[e^{\lambda(X_t - X_{t'})}\right] = e^{\lambda^2/2\mathbb{E}[X_t - X_{t'}]} = \exp\left(\frac{\lambda^2}{2} d^2(t,t')\right).$$

The following generalized process characterizes the concept of "sufficiently continuous".

**Definition 3.3.10** (Sub-Gaussian process)**.** A stochastic process $\{X_t\}_{t\in T}$ is a *sub-Gaussian process* w.r.t. $d$ if $X_t - X_s \sim \mathrm{Subg}(d^2(t,s))$. Assume $\mathbb{E}[X_t] = 0$ for all $t \in T$, then equivalently, for all $t \neq s \in T$ and $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[e^{\lambda(X_t - X_s)}\right] \leq \exp\left(\frac{\lambda^2}{2} d^2(t,s)\right).$$

It's clear that the sub-Gaussian condition encodes a strong notion of continuity (in probability) of the stochastic process $\{X_t\}_{t\in T}$ w.r.t. $d$.

**Example** (Gaussian process)**.** We see that $d(t,t') = \sqrt{\mathbb{E}\left[(X_t - X_{t'})^2\right]}$ is the naturally induced pseudo-metric such that a Gaussian process is sub-Gaussian.

**Example** (Rademacher process)**.** Consider the unnormalized Rademacher width of a set $T \subseteq \mathbb{R}^n$,

$$R_n(T) = \mathbb{E}\left[\sup_{t\in\mathbb{R}^n} \sum_{i=1}^n \epsilon_i t_i\right].$$

Let $X_t = \langle \epsilon, t\rangle$, then from Lemma 2.3.3, $X_t - X_{t'} = \langle \epsilon, t - t'\rangle \sim \mathrm{Subg}(\|t - t'\|_2^2)$, i.e., $X_t \sim \mathrm{Subg}$ w.r.t. $\|\cdot\|_2$. This is the so-called *Rademacher process*.

Inspired by the above example, one can also define the Gaussian width.

**Definition 3.3.11** (Gaussian width)**.** Let $g_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Then the *Gaussian width* of a set $A \subseteq \mathbb{R}^n$ is defined as

$$\mathrm{GW}_n(A) = \mathbb{E}\left[\sup_{a\in A} \sum_{i=1}^n \frac{1}{n} g_i a_i\right].$$

This means that the Rademacher process can be slightly modified as follows.

**Example.** If $X_t = \langle g, t\rangle$ where $g$ is a random Gaussian vector, then $X_t \sim \mathrm{Subg}$ w.r.t. $\|\cdot\|_2$.

**Theorem 3.3.2** (Gaussian width v.s. Rademacher width)**.** For any $n \geq 1$ and any set $T \subseteq \mathbb{R}^n$,

$$R_n(T) \leq \mathrm{GW}_n(T) \leq \sqrt{\log n} R_n(T).$$

Let's look at some examples of (unnormalized) Rademacher width.

**Example.** $R(B_\infty^n) = n$, $R(B_2^n) = \sqrt{n}$, and $R(B_1^n) = 1$.

**Proof.** We see that

- for $\ell_\infty$, the supremum is achieved by matching signs of $\epsilon$, which gives $R_n(B_\infty^n) = n$;
- for $\ell_2$ the supremum is achieved by choosing $t = \epsilon/\|\epsilon\|_2$, then we get $R(B_2^n) = \mathbb{E}[\epsilon] = \sqrt{n}$;
- for $\ell_1$, from Hölder's inequality, $\langle \epsilon, t\rangle \leq \|\epsilon\|_\infty \|t\|_1 = 1$.

⊛

**Example** (Supremum of empirical process). Let $\mathscr{F}$ be a class of functions bounded by 1. Let $X_f = \sqrt{n}(\mathbb{P}_n f - \mathbb{P}f)$, and consider $\{X_f\}_{f\in\mathscr{F}}$. Then,

$$X_f - X_g = \sqrt{n}\cdot\frac{1}{n}\sum_{i=1}^{n}(\underbrace{f(x_i)-g(x_i)-\mathbb{P}f-\mathbb{P}g}_{\leq 2\|f-g\|_\infty}) \sim \text{Subg}\left(4\|f-g\|_\infty^2\right),$$

hence $\{X_f\}_{f\in\mathscr{F}} \sim$ Subg w.r.t. $d(f,g) = 2\|f-g\|_\infty$.

These are all simple sets. For an arbitrary set, however, we need more general tools to compute the Rademacher width. Firstly, recall the following.

**Definition 3.3.12** (Diameter). The *diameter* of $(T,d)$ is defined as $\text{diam}(T) = \sup_{t,t'\in T} d(t,t')$.

### 3.3.4 Single-scale Bound for Expected Supremum of Sub-Gaussian Process

We're going to see the most sophisticated tools in this course. We first see a preliminary version of which and generalize it later.

**Lemma 3.3.2** (Single-scale bound). Let $\{X_t\}_{t\in T}$ be a centered sub-Gaussian process on $(T,d)$ w.r.t. $d$. Then

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \leq \inf_{\epsilon>0}\left(\mathbb{E}\left[\sup_{\substack{t,t'\in T:\\ d(t,t')\leq\epsilon}} X_t - X_{t'}\right] + \text{diam}(T)\sqrt{2\log N(T,d,\epsilon)}\right).$$

**Proof.** We first note that $\mathbb{E}\left[\sup_{t\in T} X_t\right] = \mathbb{E}\left[\sup_{t\in T} X_t - X_{t_0}\right]$ for some fixed $t_0 \in T$. Now, take an $\epsilon$-net $N$ with $\pi(t)\in N$ denotes the point such that $d(t,\pi(t)) \leq \epsilon$, then

$$\mathbb{E}\left[\sup_{t\in T} X_t - X_{t_0}\right] \leq \mathbb{E}\left[\sup_{t\in T} X_t - X_{\pi(t)}\right] + \mathbb{E}\left[\sup_{t\in T} X_{\pi(t)} - X_{t_0}\right]$$

Observe that $X_{\pi(t)} - X_{t_0} \sim \text{Subg}(\text{diam}^2(T))$, then the second term is a finite maximum such that

$$\mathbb{E}\left[\sup_{t\in T} X_{\pi(t)} - X_{t_0}\right] \leq \sqrt{2\,\text{diam}^2(T)\log N(T,d,\epsilon)} = \text{diam}(T)\sqrt{2\log N(T,d,\epsilon)}$$

from Lemma 2.3.4. By rewriting the first term, we have

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \leq \inf_{\epsilon>0}\left(\mathbb{E}\left[\sup_{\substack{t,t'\in T:\\ d(t,t')\leq\epsilon}} X_t - X_{t'}\right] + \text{diam}(T)\sqrt{2\log N(T,d,\epsilon)}\right).$$

■

**Notation** (Approximation error). The first term in the single-scale bound is the *approximation error*.

We see that the first term in the single-scale bound is still an infinite maximum, so it is not clear how to bound it. Typically, we have to do something crude here. There are some exceptions, though.

**Example.** For Rademacher processes, we have $\mathbb{E}\left[\sup_{t,t'\in T:\ \|t-t'\|\leq\delta}\langle\epsilon, t-t'\rangle\right] \leq \|\epsilon\|\delta \leq \sqrt{n}\delta$.

**Remark.** As $\epsilon$ decreases, the approximation error should get smaller, and the finite maximum increases. Therefore, when we use the single-scale bound we can then choose an optimum $\epsilon$ to minimize the sum of these two.

Let's see some applications of single-scale bound which show that the single-scale bound may not get the optimal rate.

**Example.** Consider a finite set $T = \{(0,0,\dots,0),(1,0,\dots,0),\dots,(1,1,\dots,1)\} \subseteq \mathbb{R}^n$, i.e., the footprint of the boolean function class on $\mathbb{R}$ given by $\{\mathbb{1}_{x\le\theta}\}_{\theta\in\mathbb{R}}$. By Lemma 2.3.4, $R_n(T) \le \sqrt{n\log n}$.

> **As previously seen.** We claimed that $\log n$ is superfluous.
>
> We still can't remove the $\sqrt{\log n}$: from the single-scale bound, with $\operatorname{diam}(T) = \sqrt{n}$,
>
> $$R_n(T) \le \sqrt{n}\epsilon + \sqrt{n}\sqrt{\log N(T,\|\cdot\|_2,\epsilon)}.$$

To remove $\log n$, $\epsilon$ needs to be $O(1)$ for the first term. But then $\log N(T,\|\cdot\|_2,\epsilon) \to \infty$, and we fail.

Now, let's revisit the previous example, and recall the following.

> **As previously seen.** For a class of functions bounded by 1, $X_f \sim \operatorname{Subg}(2^2\|f-g\|_\infty^2)$, i.e., $X_f - X_g \le c\sqrt{n}\|f-g\|_\infty$ almost surely.

**Example** (Empriical process supremum of $\mathcal{S}_1$)**.** Consider $X_f = \sqrt{n}(\mathbb{P}_n f - \mathbb{P}f)$ on $\mathscr{F} = \mathcal{S}_1$, i.e., functions bounded by 1 and are 1-Lipschitz on $[0,1]$. So in particular, $X_f - X_g \le c\sqrt{n}\|f-g\|_\infty$. From the single-scale bound and Theorem 3.3.1,

$$\mathbb{E}\left[\sup_{f\in\mathscr{F}} X_f\right] \le c\left(\sqrt{n}\epsilon + \sqrt{1/\epsilon}\right) = c(\sqrt{n}\cdot n^{-1/3})$$

by letting $\epsilon = n^{-1/3}$ (where this bound is minimized), giving us

$$\mathbb{E}\left[\sup_{f\in\mathcal{S}_1} \mathbb{P}_n f - \mathbb{P}f\right] \le \frac{c}{n^{1/3}}.$$

This is the first non-trivial bound we have shown besides boolean function classes.

However, observe that $X_f - X_g \le C\sqrt{n}\|f-g\|_\infty$ implies $X_f - X_g \le \|f-g\|_\infty$ in probability. The fact that we are stuck with the above almost surely bound and don't know how to incorporate this additional information, suggests that this bound is still not optimal.

> **Remark.** The optimal bound for $\mathcal{S}_1$ is $c/\sqrt{n}$, i.e., the CLT rate.

It's perhaps surprising that for the class of functions $\mathcal{S}_1$, we get the $O(n^{-1/2})$ rate for the supremum of the empirical process, because even for a single function $f \in \mathcal{S}_1$, we would still have got the $O(n^{-1/2})$ rate. This is not always the case, though. For Lipschitz function defined on $[0,1]^d$, the rates are slower. We state this result without proof for now.

**Lemma 3.3.3.** Let $S_{1,d}$ to be the set of 1-bounded 1-Lipschitz functions w.r.t. the Euclidean norm defined on $[0,1]^d$. Then there exists a universal constant $C > 0$ such that

$$\mathbb{E}\left[\sup_{f\in S_{1,d}} \mathbb{P}_n f - \mathbb{P}f\right] \le \begin{cases} Cn^{-1/2}, & \text{if } d = 1; \\ Cn^{-1/2}\log n, & \text{if } d = 2; \\ Cn^{-1/d}\log n, & \text{if } d > 2. \end{cases}$$

These rates are tight and corresponding lower-bounds are also known [Han16, Problem 5.11 (d)].

# Lecture 12: Chaining Method and Dudley's Entropy Bound

20 Sep. 9:00

### 3.3.5 Dudley's Entropy Bound

To overcome the limitation of the single-scale bound, we can repeatedly take $\epsilon$-net, which is considered as a *multi-scale bound*. The theorem requires one technical assumption of the stochastic process.

**Definition 3.3.13** (Separable). We say that $\{X_t\}_{t\in T}$ is a *separable* process if there exists a countable $T_0 \subseteq T$ such that (outside a null set) for all $t \in T$, there exists $\{t_n \in T_0\}_n$ such that $d(t_n, t) \to 0$ satisfying $\lim_{n\to\infty} X_{t_n} = X_t$.

It's clear that $\sup_{t\in T_0} X_t = \sup_{t\in T} X_t$. Moreover, this notion is consistent with the separability of a topological space[2] we saw in real analysis.

**Example** (Separable metric space). If $(T, d)$ is separable (as a topological space), $\{X_t\}$ has countable sample path almost surely, then $\{X_t\}$ is separable.

Now, we can state the bound we want.

**Theorem 3.3.3** (Dudley's entropy bound). Let $\{X_t\}_{t\in T}$ be a centered and separable sub-Gaussian process on $(T, d)$ w.r.t. $d$. Then

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \le 6\sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\log N(T, d, 2^{-k})}.$$

**Proof.** Consider the case that $|T| < \infty$ and $|T| = \infty$.

**Claim.** The result holds for $|T| < \infty$.

**Proof.** Let $K_0$ be the largest integer such that $2^{-K_0} \ge \operatorname{diam}(T)$, and let $K_1$ be the smallest integer such that $0 < 2^{-K_1} < \min_{s\neq t\in T} d(s, t)$. Then we let $N_k$ be a $2^{-k}$-net of $T$ such that

- $k = K_0$: $N_{K_0} = \{t_0\}$ is a $2^{-K_0}$-net of $T$ for a fixed $t_0 \in T$.
- $k = K_1$: $N_{K_1} = T$ is a $2^{-K_1}$-net of $T$.

Write $\pi_k(t)$ for the closest element in $N_k$ to $t$, in particular, $d(t, \pi_k(t)) \le 2^{-k}$. By writing

$$X_t - X_{t_0} = X_{\pi_{K_1}(t)} - X_{\pi_{K_0}(t)} = X_{\pi_{K_1}(t)} - X_{\pi_{K_1-1}(t)} + X_{\pi_{K_1-1}(t)} - \cdots + X_{\pi_{K_0+1}(t)} - X_{\pi_{K_0}(t)},$$

hence we have

$$\begin{aligned}\mathbb{E}\left[\sup_{t\in T} X_t\right] &= \mathbb{E}\left[\sup_{t\in T} X_t - X_{t_0}\right] \\ &= \mathbb{E}\left[\sup_{t\in T}\sum_{k=K_0+1}^{K_1}\left(X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\right)\right] \le \sum_{k=K_0+1}^{K_1}\mathbb{E}\left[\sup_{t\in T}\left(X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\right)\right].\end{aligned}$$

Since the cardinality of $\{X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\}_{t\in T}$ is $|N_k||N_{k-1}| \le |N_k|^2$, and

$$X_{\pi_k(t)} - X_{\pi_{k-1}(t)} \sim \operatorname{Subg}(d^2(\pi_k(t), \pi_{k-1}(t)))$$

with $d(\pi_k(t), \pi_{k-1}(t)) \le d(\pi_k(t), t) + d(t, \pi_{k-1}(t)) \le 2^{-k} + 2^{-k+1} \le 3\cdot 2^{-k}$, from Lemma 2.3.4,

$$\mathbb{E}\left[\sup_{t\in T}\left(X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\right)\right] \le 3\times 2^{-k}\sqrt{2\log|N_k|^2} = 6\times 2^{-k}\sqrt{\log|N_k|}$$

for each $k$. Summing over $k$ yields the result. ⊛

[2] A topological space is *separable* if it contains a countable dense subset.

**Claim.** The result holds for $|T| = \infty$.

**Proof.** From separability, there exists a countable $T_0$ such that $\mathbb{E}\left[\sup_{t\in T_0} X_t\right] = \mathbb{E}\left[\sup_{t\in T} X_t\right]$. Let $T_k$ be a countable approximation of $T_0$, then $\sup_{t\in T_k} X_t \to \sup_{t\in T_0} X_t$ as $k \to \infty$, so

$$\mathbb{E}\left[\sup_{t\in T_k} X_t\right] \to \mathbb{E}\left[\sup_{t\in T_0} X_t\right] = \mathbb{E}\left[\sup_{t\in T} X_t\right] \text{ as } k \to \infty$$

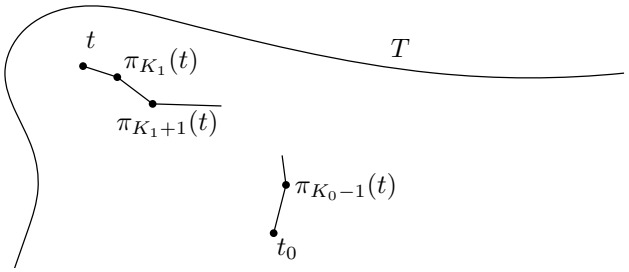
from the monotone convergence theorem. Hence, it suffices to bound $\mathbb{E}\left[\sup_{t\in T_k} X_t\right]$ instead of $\mathbb{E}\left[\sup_{t\in T} X_t\right]$ for each $k$. As $|T_k| < \infty$ and $N(T_k, d, 2^{-k}) \le N(T_0, d, 2^{-k})$ for all $k$,

$$6\sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\log N(T_k, d, 2^{-k})} \le 6\sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\log N(T_0, d, 2^{-k})}.$$

⊛

■

**Note** (Chaining method). This method is called *chaining* since we're constructing a chain of $X_{\pi_k(t)}$, with smaller and smaller distances.




An alternative integral form of Dudley's entropy bound is given by the following.

**Corollary 3.3.1** (Dudley integral entropy bound). Let $\{X_t\}_{t\in T}$ be a centered and separable sub-Gaussian process on $(T, d)$ w.r.t. $d$. Then

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \le 12\int_0^{\operatorname{diam}(T)} \sqrt{\log N(T, d, \epsilon)}\,\mathrm{d}\epsilon.$$

**Proof.** Observe that

$$\begin{aligned}
\sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\log N(T, d, 2^{-k})} &= 2\sum_{k\in\mathbb{Z}} \int_{2^{-k-1}}^{2^{-k}} \sqrt{\log N(T, d, 2^{-k})}\,\mathrm{d}\epsilon \\
&\le 2\sum_{k\in\mathbb{Z}} \int_{2^{-k-1}}^{2^{-k}} \sqrt{\log N(T, d, \epsilon)}\,\mathrm{d}\epsilon && N(T,d,\epsilon) \nearrow \text{ as } \epsilon \searrow \\
&= 2\int_0^{\infty} \sqrt{\log N(T, d, \epsilon)}\,\mathrm{d}\epsilon \\
&= 2\int_0^{\operatorname{diam}(T)} \sqrt{\log N(T, d, \epsilon)}\,\mathrm{d}\epsilon. && \epsilon > \operatorname{diam}(T),\ N(T,d,\epsilon) = 1
\end{aligned}$$

■

Now, we note that we have finally reached the optimal bound for $\mathcal{S}_1$, solving the problems we saw in the previous example.

**Example** (Supremum of empirical process for $\mathcal{S}_1$). Consider the separable sub-Gaussian process $X_f = \sqrt{n}(\mathbb{P}_n f - \mathbb{P}f)$ for $\mathscr{F} = \mathcal{S}_1$. In particular, $f, g \in \mathscr{F}$ are 1-Lipschitz on $[0,1]$ satisfying $|f|, |g| \le 1$ and $X_f - X_g \in \mathrm{Subg}(2^2\|f-g\|_\infty^2)$. Since $\mathrm{diam}(\mathscr{F}) = 2$ and for all $\epsilon < 1/2$,

$$N(\mathcal{S}_1, \|\cdot\|_\infty, \epsilon) = \exp(c/\epsilon)$$

from Theorem 3.3.1. Then by the Dudley's integral entropy bound,

$$\mathbb{E}\left[\sup_{f\in\mathscr{F}} X_f\right] \le 12\int_0^2 \sqrt{\log N(\mathcal{S}_1, \|\cdot\|_\infty, \epsilon)}\,\mathrm{d}\epsilon = 12\int_0^2 \sqrt{\frac{c}{\epsilon}}\,\mathrm{d}\epsilon = 24\sqrt{2c} < O_n(1).$$

Dividing both sides by $\sqrt{n}$, we achieve the optimal rate $\mathbb{E}\left[\sup_{f\in\mathscr{F}}(\mathbb{P}_n f - \mathbb{P}f)\right] = O(1/\sqrt{n})$.

**Remark.** The Dudley's integral entropy bound for $\mathcal{S}_\alpha$ is also finite; while for function classes with covering number as $\exp(c/\epsilon^2)$ is divergent.

# Lecture 13: More on Chaining

Let's see some alternate forms of Dudley's integral entropy bound. In the following, assume that $\{X_t\}_{t\in T}$ is a centered and separable sub-Gaussian process on $(T,d)$ w.r.t. $d$. 22 Sep. 9:00

**Corollary 3.3.2** (Difference form). The same bound as the Dudley's integral entropy bound holds for $\mathbb{E}\left[\sup_{t\in T}|X_t - X_{t_0}|\right]$ and $\mathbb{E}\left[\sup_{s,t\in T}|X_s - X_t|\right]$.

**Proof.** The former is clear, and the latter can be proved by triangle inequality with $X_s - X_{t_0}$. ■

**Corollary 3.3.3** (High probability form). The high probability bound version holds:

$$\mathbb{P}\left(\sup_{s,t\in T}|X_s - X_t| \le C\left(\int_0^\infty \sqrt{\log N(T,d,\epsilon)}\,\mathrm{d}\epsilon + u\,\mathrm{diam}(T)\right)\right) \ge 1 - 2e^{-u^2}.$$

**Proof.** See Corollary A.2.1 for a proof. ■

**Corollary 3.3.4** (Finite resolution form). The following generalizes the Dudley's integral entropy bound in the sense that $\delta > 0$:

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \le C\left(\mathbb{E}\left[\sup_{\substack{t,t'\in T\\ d(t,t')\le\delta}} X_t - X_{t'}\right] + \int_\delta^\infty \sqrt{\log N(T,d,\epsilon)}\,\mathrm{d}\epsilon\right).$$

**Proof.** See Corollary A.2.2 for a proof. ■

The finite resolution version is useful since the entropy, integral can diverge, e.g., if $\log N(t,d,\epsilon) = \Omega(1/\epsilon^2)$. Moreover, this can be used to show Lemma 3.3.3.

**Remark.** We can moreover write

$$\begin{aligned}\mathbb{E}\left[\sup_{t\in T} X_t\right] &\le C\int_0^{\mathrm{diam}(T)} \sqrt{\log N(T,d,\epsilon)}\,\mathrm{d}\epsilon\\ &\le C\left(\int_0^{\mathrm{diam}(T)/2} \sqrt{\log N(T,d,\epsilon)}\,\mathrm{d}\epsilon + \int_{\mathrm{diam}(T)/2}^{\mathrm{diam}(T)} \sqrt{\log N(T,d,\epsilon)}\,\mathrm{d}\epsilon\right)\\ &\le 2C\int_0^{\mathrm{diam}(T)/2} \sqrt{\log N(T,d,\epsilon)}\,\mathrm{d}\epsilon.\end{aligned}$$

### 3.3.6 Uniform Entropy Integral Bound

Let's discuss some limitations of the Dudley's integral entropy bound. First, recall the following.

> **As previously seen.** In the example of the optimal rate for $\mathcal{S}_1$, whenever
>
> $$\int_0^\infty \sqrt{\log N(T,d,\epsilon)}\,\mathrm{d}\epsilon < \infty \Rightarrow \mathbb{E}\left[\sup_f \mathbb{P}_n f - \mathbb{P}f\right] \le c/\sqrt{n}.$$

Note that we're doing chaining w.r.t. $\|\cdot\|_\infty$ on $\mathscr{F}$ so far. To see its limitation, consider again the boolean function classes $\mathscr{F}$ and let $X_f = \sqrt{n}(\mathbb{P}_n f - \mathbb{P}f) \sim \mathrm{Subg}(c^2\|\cdot\|_\infty^2)$. From Proposition 3.2.1,

$$\mathbb{E}\left[\sup_{f\in\mathscr{F}} \mathbb{P}_n f - \mathbb{P}f\right] \le \sqrt{\frac{\mathrm{VC}(\mathscr{F})\log n}{n}}.$$

Now, observe that for any $f \neq g$ in $\mathscr{F}$, $\|f-g\|_\infty = 1$. This implies that by taking $\epsilon \in (0,1)$,

$$N(\mathscr{F}, \|\cdot\|_\infty, \epsilon) = |\mathscr{F}| = \infty$$

for any interesting case, e.g., $\mathscr{F} = \{\mathbb{1}_{x\le\theta}\}_{\theta\in\mathbb{R}}$, i.e., chaining w.r.t. $\|\cdot\|_\infty$ only gives a vacuous bound.

> **Intuition.** To fix this, we can use the idea of the symmetrization.

Firstly, given some observed i.i.d. data $x_1,\dots,x_n$, recall the following.

> **As previously seen.** By conditioning on the data $x_1,\dots,x_n$, symmetrization shows that
>
> $$\mathbb{E}\left[\sup_{f\in\mathscr{F}} \sqrt{n}(\mathbb{P}_n f - \mathbb{P}f)\right] \le \frac{2}{\sqrt{n}} R_n(\mathscr{F}) = 2\mathbb{E}_{x,\epsilon}\left[\sup_{f\in\mathscr{F}} \frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f(x_i)\right].$$

Specifically, we want to look at $\mathbb{E}_x\left[\mathbb{E}_\epsilon\left[\sup_{f\in\mathscr{F}} X_f\right]\right]$ and compute the Rademacher width. Let $X_f = \frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f(x_i)$, we have

$$X_f - X_g = \frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i\big(f(x_i)-g(x_i)\big) \sim \mathrm{Subg}\big(\big\|\big(f(x_i)\big)_i - \big(g(x_i)\big)_i\big\|_2^2\big) = \mathrm{Subg}\left(\frac{1}{n}\sum_{i=1}^n \big(f(x_i)-g(x_i)\big)^2\right),$$

where $(f(x_i))_i = (f(x_1),\dots,f(x_n))$. Hence, $X_f \sim \mathrm{Subg}\left(\frac{1}{n}\sum_i (f(x_i)-g(x_i))^2\right)$.

> **Note.** We're already doing better since $\sqrt{\frac{1}{n}\sum_i (f(x_i)-g(x_i))^2} \le \|f-g\|_\infty$.

We see that $\sqrt{\frac{1}{n}\sum_i (f(x_i)-g(x_i))^2}$ is similar to $\|f-g\|_2$, but just on the empirical measure (with i.i.d. data $x_i$'s). Hence, consider the following notation.

> **Notation.** Let $L_2(\mathbb{P}_n)$ denote the metric w.r.t. $\mathbb{P}_n$[a] such that
>
> $$L_2^2(\mathbb{P}_n)(f,g) \coloneqq \frac{1}{n}\sum_{i=1}^n \big(f(x_i)-g(x_i)\big)^2.$$
>
> [a] Formally, $\mathbb{P}_n$ is the empirical measure uniform on $\{x_i\}_{i=1}^n$.

In our new notation, $X_f \sim \mathrm{Subg}(L_2(\mathbb{P}_n))$. Now, we can do the chaining argument on $L_2(\mathbb{P}_n)$ and get

$$\mathbb{E}\left[\sup \sqrt{n}(\mathbb{P}_n f - \mathbb{P}f)\right] \le C\int_0^{\mathrm{diam}(\mathscr{F})} \sqrt{\log N(\mathscr{F}, L_2(\mathbb{P}_n), \epsilon)}\,\mathrm{d}\epsilon,$$

where

$$\mathrm{diam}(\mathscr{F}) = \sup_{f,g} L_2(\mathbb{P}_n)(f,g) = \frac{1}{n}\sum_{i=1}^n \big(f(x_i)-g(x_i)\big)^2 \le \sup_{f\in\mathscr{F}} \sqrt{\mathbb{P}_n f^2},$$

hence we have

$$\mathbb{E}\left[\sup \sqrt{n}(\mathbb{P}_n f - \mathbb{P}f)\right] \le C \cdot \mathbb{E}_x \left[\int_0^{\sup_{f\in\mathscr{F}} \sqrt{\mathbb{P}_n f^2}} \sqrt{\log N(\mathscr{F}, L_2(\mathbb{P}_n), \epsilon)}\, \mathrm{d}\epsilon\right].$$

However, there's a problem.

**Problem.** $L_2(\mathbb{P}_n)$ is a "random" metric, so $N(\mathscr{F}, L_2(\mathbb{P}_n), \epsilon)$ is hard to compute.

**Answer.** To resolve this, we take the supremum over all measures $\mu$ supported on $\chi$, i.e.,

$$C\mathbb{E}_x \left[\int_0^{\sup_{f\in\mathscr{F}} \sqrt{\mathbb{P}_n f^2}} \sqrt{\log N(\mathscr{F}, L_2(\mathbb{P}_n), \epsilon)}\, \mathrm{d}\epsilon\right] \le C\mathbb{E}_x \left[\int_0^{\sup_{f\in\mathscr{F}} \sqrt{\mathbb{P}_n f^2}} \sqrt{\sup_\mu \log N(\mathscr{F}, L_2(\mu), \epsilon)}\, \mathrm{d}\epsilon\right].$$

⊛

This might seem very bad, but actually it's not since $L_2(\mu) < L_\infty$. Specifically, to bound this supremum over all measures, consider the following.

**Definition 3.3.14** (Koltchinskii-Pollard entropy)**.** The *Koltchinskii-Pollard entropy* of $\mathscr{F}$ is defined as

$$\sup_\mu \log N(\mathscr{F}, L_2(\mu), \epsilon).$$

**Example.** For boolean function classes, $\sup_f \sqrt{\mathbb{P}_n f^2} \le 1$.

We then have the following for the boolean function classes.

**Intuition** (Main bound)**.** Let $\mathscr{F}$ be a boolean function class, then since $\sup_{f\in\mathscr{F}} \sqrt{\mathbb{P}_n f^2} \le 1$,

$$\mathbb{E}\left[\sup_{f\in\mathscr{F}} \sqrt{n}|\mathbb{P}_n f - \mathbb{P}f|\right] \le C\mathbb{E}_x \left[\int_0^1 \sqrt{\sup_\mu \log N(\mathscr{F}, L_2(\mu), \epsilon)}\, \mathrm{d}\epsilon\right].$$

More generally, if we have $F \ge f$ (called envelope) for all $f \in \mathscr{F}$ such that $\mathbb{P}F^2 < \infty$, this holds.

**Problem.** How can we compute the Koltchinskii-Pollard entropy?

**Answer.** We can use some notions of combinatorial dimension (e.g., VC dimension) upper-bounds the Koltchinskii-Pollard entropy such that

$$\sup_\mu N(\mathscr{F}, L_2(\mu), \epsilon) \le (c_1/\epsilon)^{c_2 \times \mathrm{VC}(\mathscr{F})} \approx \epsilon^{-d}$$

for $d$ being "dimension" (parametric). ⊛

**Remark.** This implies a $\sqrt{\mathrm{VC}(\mathscr{F})/n}$ rate (without a log term!) for $\mathbb{E}\left[\sup(\mathbb{P}_n f - \mathbb{P}f)\right]$.

# Lecture 14: Uniform Entropy Integral Bound

25 Sep. 9:00

**As previously seen.** Motivated by the fact that boolean function classes are not totally bounded w.r.t. $\ell_\infty$, $\|f - g\|_\infty = 1$, we're trying to establish the bound on

$$\mathbb{E}\left[\sup_{f\in\mathscr{F}} \sqrt{n}(\mathbb{P}_n f - \mathbb{P}f)\right] \le 2\mathbb{E}_x \left[\mathbb{E}_\epsilon \left[\sup_{f\in\mathscr{F}} \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i f(x_i)\right]\right] = 2\sqrt{n} R_n(\mathscr{F})$$

where the inner expectation is just the (scaled) Rademacher width $\sqrt{n}R_n(\{f(x_1),\dots,f(x_n)\}_{f\in\mathscr{F}})$.[a] Let $X_f = \langle \epsilon, f\rangle/\sqrt{n}$, then $\{X_f\}_{f\in\mathscr{F}}$ is sub-Gaussian w.r.t. $L_2(\mathbb{P}_n)$.[b]

[a] We can also abuse the notation $R_n(\mathscr{F})$ to replace $R_n(\{\sqrt{n}f(x_1),\dots,\sqrt{n}f(x_n)\}_{f\in\mathscr{F}})$, but let's not do this.
[b] Recall that $L_2^2(\mathbb{P}_n)(f,g) = \frac{1}{n}\sum_{i=1}^n \left(f(X_i)-g(X_i)\right)^2$, compared to $L_2(\mathbb{P})(f,g) = \int \left(f(x)-g(x)\right)^2 \mathrm{d}\mathbb{P}$.

The obstacle we're facing is the lack of control of $\sup_{f\in\mathscr{F}}\sqrt{\mathbb{P}f^2}$. Consider the following notion.

**Definition 3.3.15** (Envelope). A non-negative valued function $F\colon \chi \to [0,\infty]$ is an *envelope* for $\mathscr{F}$ if $\sup_{f\in\mathscr{F}}|f(x)| \le F(x)$ for all $x\in\chi$.

**Example.** For boolean function classes, $F(x)=1$ is an envelope.

**Remark.** Let $F$ be an envelope of $\mathscr{F}$, then $\sup_{f\in\mathscr{F}}\sqrt{\mathbb{P}_n f^2} \le \sqrt{\mathbb{P}_n F^2}$, as we want.

With this new notion, we can state the main bound we want, i.e., the uniform entropy integral bound.

**Theorem 3.3.4** (Uniform entropy integral bound). Given a function class $\mathscr{F}$ and an envelope $F$ of $\mathscr{F}$ such that $\mathbb{P}F^2 < \infty$, then for $x_1,\dots,x_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$,

$$\mathbb{E}\left[\sup_{f\in\mathscr{F}}\sqrt{n}|\mathbb{P}_n f - \mathbb{P}f|\right] \le 2\sqrt{n}R_n(\mathscr{F}) \le C\|F\|_{L_2(\mathbb{P})}\int_0^1 \sqrt{\log\sup_\mu N(\mathscr{F},L_2(\mu),\epsilon\sqrt{\mu F^2})}\,\mathrm{d}\epsilon.$$

**Proof.** Summarizing what we have established, we have

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathscr{F}}\sqrt{n}(\mathbb{P}_n f-\mathbb{P}f)\right] &\le 2\mathbb{E}_x\left[\mathbb{E}_\epsilon\left[\sup_{f\in\mathscr{F}}\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f(x_i)\right]\right] && \text{symmetrization}\\
&= 2\sqrt{n}R_n(\mathscr{F})\\
&\le C\mathbb{E}_x\left[\int_0^{\sup_{f,g\in\mathscr{F}}\frac{L_2(\mathbb{P}_n)(f,g)}{2}}\sqrt{\log N(\mathscr{F},L_2(\mathbb{P}_n),\epsilon)}\,\mathrm{d}\epsilon\right] && \text{Dudley's bound}\\
&\le C\mathbb{E}_x\left[\int_0^{\sup_{f\in\mathscr{F}}\sqrt{\mathbb{P}_n f^2}}\sqrt{\log N(\mathscr{F},L_2(\mathbb{P}_n),\epsilon)}\,\mathrm{d}\epsilon\right]\\
&\le C\mathbb{E}_x\left[\int_0^{\sqrt{\mathbb{P}_n F^2}}\sqrt{\log N(\mathscr{F},L_2(\mathbb{P}_n),\epsilon)}\,\mathrm{d}\epsilon\right]\\
&= C\mathbb{E}_x\left[\sqrt{\mathbb{P}_n F^2}\int_0^1\sqrt{\log N(\mathscr{F},L_2(\mathbb{P}_n),\epsilon\sqrt{\mathbb{P}_n F^2})}\,\mathrm{d}\epsilon\right] && \epsilon\leftarrow\sqrt{\mathbb{P}_n F^2}\epsilon\\
&\le C\mathbb{E}_x\left[\sqrt{\mathbb{P}_n F^2}\int_0^1\sqrt{\sup_\mu\log N(\mathscr{F},L_2(\mu),\epsilon\sqrt{\mu F^2})}\,\mathrm{d}\epsilon\right]\\
&\le C\left[\int_0^1\sqrt{\sup_\mu\log N(\mathscr{F},L_2(\mu),\epsilon\sqrt{\mu F^2})}\,\mathrm{d}\epsilon\right]\mathbb{E}_x\left[\sqrt{\mathbb{P}_n F^2}\right]
\end{aligned}$$

from Jensen's inequality, $\mathbb{E}\left[\sqrt{\mathbb{P}_n F^2}\right] \le \sqrt{\mathbb{E}\left[\mathbb{P}_n F^2\right]} = \sqrt{\mathbb{P}F^2} = \|F\|_{L_2(\mathbb{P})}$,

$$\le C\|F\|_{L_2(\mathbb{P})}\left[\int_0^1\sqrt{\sup_\mu\log N(\mathscr{F},L_2(\mu),\epsilon\sqrt{\mu F^2})}\,\mathrm{d}\epsilon\right].$$

■

**Notation.** We sometimes denote $\int_0^1\sqrt{\log\sup_\mu N(\mathscr{F},L_2(\mu),\epsilon\sqrt{\mu F^2})}\,\mathrm{d}\epsilon$ by $\boldsymbol{J}(F,\mathscr{F})$.

$\mathscr{F}$ needs not to be bounded, instead, what we really need is an envelope:

**Remark.** If we apply the above bound to a bounded function class then we do not need the notion of an envelope. The assumption of an envelope is slightly more general than assuming boundedness.

**Remark.** The Koltchinskii-Pollard entropy integral is free from $\mathbb{P}$, while $\|F\|_{L_2(\mathbb{P})}$ depends on $\mathbb{P}$. Thus, the overall bound is distribution-free, as has been all of our bounds so far, if the function class is bounded.

**Example.** For boolean function classes, $\|F\|_{L_2(\mathbb{P})} \leq 1$, so the bound is uniform over $\mathbb{P}$.

### 3.3.7 Uniform $L_2$ Entropy is Bounded by Combinatorial Dimension

We're now ready to revisit the problem we asked in the previous lecture:

**Problem.** How can be bound the uniform $L_2$-entropy, i.e., Koltchinskii-Pollard entropy?

**Answer.** For boolean function classes, Koltchinskii-Pollard entropy can be bounded in terms of VC dimension; for non-boolean function classes, an extended notion of VC dimension is needed. ⊛

**Theorem 3.3.5** (Dudley). Let $\mathscr{F}$ be a boolean function class, then there exist absolute constants $c_1, c_2$ such that for all $0 < \epsilon < 1$,

$$\sup_{\mu} N(\mathscr{F}, L_2(\mu), \epsilon) \leq \left(\frac{c_1}{\epsilon}\right)^{c_2 \operatorname{VC}(\mathscr{F})}$$

**Proof.** It suffices to bound the packing number instead of the covering number from Lemma 3.3.1. To upper-bound the packing number via $d \coloneqq \operatorname{VC}(\mathscr{F})$, first fix a probability measure $\mu$ on $\chi$, consider a maximum $\epsilon$-packing $M = \{f_1, \ldots, f_N\}$ of $\mathscr{F}$ w.r.t. $L_2(\mu)$. Then for all $i \neq j$,

$$\int (f_i - f_j)^2 \, \mathrm{d}\mu = \mu(f_i \neq f_j) > \epsilon^2.$$

Then, sample $K$ points $W_1, \ldots, W_K$ i.i.d. from $\mu$, we want that all $\{f_i\}_{i=1}^N$ to have different values on $(w_1, \ldots, w_K)$. Note that for $i \neq j$,

$$\mu(f_i = f_j \text{ on } w_1, \ldots, w_K) \leq (1 - \epsilon^2)^K \leq e^{-K\epsilon^2}$$

from $(1 - x)^k \leq e^{-kx}$. This implies

$$\mu(\exists \text{ at least one pair } i \neq j \text{ such that } f_i = f_j \text{ on } w_1, \ldots, w_K) \leq \binom{N}{2} e^{-K\epsilon^2},$$

hence

$$\mu(\text{all the } f_i\text{'s are distinct on } w_1, \ldots, w_k\ ) \geq 1 - \binom{N}{2} e^{-K\epsilon^2} \geq \frac{1}{2}$$

by choosing $K\epsilon^2 \approx 2 \log N$. We conclude that there exists $K$ points $w_1, \ldots, w_K$ such that all the $f_i$'s are distinct on $\{w_1, \ldots, w_K\}$. From Sauer-Shelah lemma,[a]

$$N = |\mathscr{F}(w_1, \ldots, w_K)| \leq \left(\frac{eK}{d}\right)^d = \left(\frac{2e \log N}{\epsilon^2 d}\right)^d.$$

We see that $N \leq (\log N)^d$. To further bound $N$, consider[b]

$$N^{1/d} \leq \frac{4e \log N}{2d\epsilon^2} = \frac{4e}{\epsilon^2} \log N^{1/2d} \leq \frac{4e}{\epsilon^2} N^{1/2d}$$

from $\log x \le x$, hence $N^{1/2d} \le 4e/\epsilon^2$, or equivalently,

$$N \le (4e)^{2d} \left(\frac{1}{\epsilon}\right)^{4d} = \left(\frac{2\sqrt{e}}{\epsilon}\right)^{4d} =: \left(\frac{c_1}{\epsilon}\right)^{c_2 d}.$$

■

[a] Note that we have only shown the case for $K \ge \mathrm{VC}(\mathscr{F})$. However, for $K < \mathrm{VC}(\mathscr{F})$, it's also easy to show.
[b] We want to make the exponent $a$ of $\log N^a$ to be less than $1/d$.

**Remark.** For boolean function classes, while $N(\mathscr{F}, L_\infty, \epsilon) = \infty$, the above shows that

$$\sup_\mu N(\mathscr{F}, L_2(\mu), \epsilon) < \infty.$$

We can now finally generalize Proposition 3.2.1. Recall the following.

**As previously seen.** From Proposition 3.2.1, for boolean function classes, we have

$$\mathbb{E}\left[\sup_{f\in\mathscr{F}} \sqrt{n}|\mathbb{P}_n f - \mathbb{P} f|\right] \le c\sqrt{\mathrm{VC}(\mathscr{F}) \log \frac{en}{\mathrm{VC}(\mathscr{F})}}.$$

**Corollary 3.3.5.** Let $\mathscr{F}$ be a boolean function class, for some constant $C$, we have

$$\mathbb{E}\left[\sup_{f\in\mathscr{F}} \sqrt{n}|\mathbb{P}_n f - \mathbb{P} f|\right] \le C\sqrt{\mathrm{VC}(\mathscr{F})}.$$

**Proof.** Applying uniform entropy integral bound, with $\|F\|_{L_2(\mathbb{P})} = 1$ and Theorem 3.3.5,

$$\mathbb{E}\left[\sup_{f\in\mathscr{F}} \sqrt{n}|\mathbb{P}_n f - \mathbb{P} f|\right] \le C\int_0^1 \sqrt{c_2 \,\mathrm{VC}(\mathscr{F}) \log \frac{c_1}{\epsilon}}\, \mathrm{d}\epsilon \le C'\sqrt{\mathrm{VC}(\mathscr{F})}\int_0^1 \log \frac{1}{\epsilon}\, \mathrm{d}\epsilon \le C'\sqrt{\mathrm{VC}(\mathscr{F})}.$$

■

**Remark.** Compare Proposition 3.2.1 to Corollary 3.3.5, the extra $\sqrt{\log n}$ factor in the bound which we have now got rid of thanks to chaining. The bound holds for Rademacher complexity and as a consequence for suprema of empirical process.

**Note.** Consider the classification problem in the statistical learning setting, where $\hat{f}$ is the ERM over a given boolean function class. Then the excess risk is also bounded by $C\sqrt{\mathrm{VC}(\mathscr{F})/n}$.

**Proof.** From symmetrization, the excess risk can be bounded as

$$\mathbb{E}\left[L(\hat{f})\right] - \inf_{f\in\mathscr{F}} \mathbb{E}\left[L(\hat{f})\right] \le \mathbb{E}\left[\sup_{f\in\mathscr{F}} \left(\mathbb{E}\left[f(X)\right] - \frac{1}{n}\sum_{i=1}^n f(x_i)\right)\right] \le 2R_n(\mathscr{F}).$$

From Corollary 3.3.5, we finally show that in this example, the $\sqrt{\log n}$ factor is superfluous as well. ⊛

A final remark is the following.

**Remark.** Corollary 3.3.5 is a distribution-free result, i.e.,

$$\sup_{\mathbb{P}} \mathbb{E}\left[\sup_{f\in\mathscr{F}} |\mathbb{P}_n f - \mathbb{P} f|\right] \le C\sqrt{\frac{\mathrm{VC}(\mathscr{F})}{n}}.$$

This means that boolean function classes with finite VC dimension are uniform Glivenko-Cantelli.

We have been looking at non-boolean function classes, where the "machinery" we have done is the following:

1. Bound uniform entropy w.r.t. $L_2$ (uniform entropy integral bound).
2. Uniform $L_2$ entropy $\leq$ VC dimension (Theorem 3.3.5).

**Problem.** How to extend this "machinery" to non-boolean function classes?

**Answer.** Define VC dimension for non-boolean function classes. ⊛

# Lecture 15: Parametric v.s. Non-Parametric

## 3.3.8 Parametric versus Non-Parametric Function Classes

27 Sep. 9:00

We start by asking the following question.

**Problem.** What makes a function class "parametric" or "non-parametric"?

**Answer.** If the function class is a vector space, we can usually use the linear algebra notion of dimension. ⊛

Consider the following (not very precise) definition in terms of the uniform $L_2$ entropy.

**Definition 3.3.16** (Parametric). A function class $\mathscr{F}$ is *parametric* if there exists a notion of dimension $\dim \mathscr{F}$ and a constant $C$ such that

$$\sup_{\mu} N(\mathscr{F}, L_2(\mu), \epsilon) \leq \left(\frac{C}{\epsilon}\right)^{\dim(\mathscr{F})}.$$

**Example.** Boolean function class on $\chi$ with finite VC dimension is parametric.

**Proof.** Dudley's result directly applies. ⊛

**Definition 3.3.17** (Non-parametric). A function class $\mathscr{F}$ is *non-parametric* if there is a $p > 0$ and a constant $C$ such that

$$\sup_{\mu} \log N(\mathscr{F}, L_2(\mu), \epsilon) \leq \left(\frac{C}{\epsilon}\right)^{p}.$$

Let's consider any parametric class $\mathscr{F}$ uniformly bounded by 1 with $\dim \mathscr{F} = d$. Then from Dudley's theorem, the Rademacher complexity can be bounded as

$$\mathbb{E}_x\left[\mathbb{E}_\epsilon\left[\sup_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(x_i)\right]\right] \leq \frac{12}{\sqrt{n}} \int_0^1 \sqrt{\sup_{\mu} \log N(\mathscr{F}, L_2(\mu), \epsilon)}\, \mathrm{d}\epsilon \leq \frac{12}{\sqrt{n}} \int_0^1 \sqrt{d \log \frac{C}{\epsilon}}\, \mathrm{d}\epsilon \leq C' \sqrt{\frac{d}{n}}.$$

Hence, we get the parametric rate $O(\sqrt{d/n})$, and this result is distribution-free.

Analogously, we want to know what we will get for a non-parametric function class (uniform bounded

by 1)? Now, since for a non-parametric class, the uniform $L_2$ entropy is $\leq (C/\epsilon)^p$,

$$\begin{aligned}
&\mathbb{E}_\epsilon\left[\sup_{f\in\mathscr{F}}\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f(x_i)\right] && \frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f(x_i) = X_f \sim \mathrm{Subg}(L_2(\mathbb{P}_n)) \\
&\leq \mathbb{E}\left[\sup_{\substack{f,g\in\mathscr{F}:\\ L_2(\mathbb{P}_n)(f,g)\leq\delta}} X_f - X_g\right] + \int_\delta^1 \sqrt{\sup_\mu \log N(\mathscr{F}, L_2(\mu), \epsilon)}\,\mathrm{d}\epsilon && \text{modified Corollary 3.3.4} \\
&\leq \sqrt{n}\cdot\delta + \int_\delta^1 \left(\frac{C}{\epsilon}\right)^{p/2}\mathrm{d}\epsilon
\end{aligned}$$

Since the choice of $\delta$ is arbitrary, we can optimize in terms of $\delta$. There are three cases:

- $p < 2$: Take $\delta = 0$ because the integral converges, and we get a parametric rate bound for $R_n$ with some constant $c$:
$$\mathbb{E}\left[\sup_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)\right] \leq \frac{c}{\sqrt{n}}.$$

> **Remark.** $O(1/\sqrt{n})$ is a parametric rate.

> **Example.** Consider Hölder smooth classes. Even though these function classes are non-parametric according to Definition 3.3.17, in terms of $R_n$ or supremum of empirical process, the rate is still parametric.

- $p > 2$: We see that for all $\delta \in (0, 1)$,
$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)\right] &\leq \delta + \frac{1}{\sqrt{n}}\int_\delta^1 \left(\frac{C}{\epsilon}\right)^{p/2}\mathrm{d}\epsilon \\
&= \delta + \frac{C^{p/2}}{\sqrt{n}}\cdot\frac{\epsilon^{-p/2+1}}{1-p/2}\Big|_\delta^1 \\
&\approx \delta + \frac{1}{\sqrt{n}}\left(-\epsilon^{-p/2+1}\right)\Big|_\delta^1 && \text{dropping constant } (1-p/2<0) \\
&\approx \delta + \frac{1}{\sqrt{n}}\delta^{1-p/2}.
\end{aligned}$$

  Now by optimizing over $\delta$, setting $\delta = \delta^{1-p/2}/\sqrt{n}$, we get the bound $O(n^{-1/p})$.

> **Remark.** $O(n^{-1/p})$ is a non-parametric rate, strictly slower than the parametric rate $O(n^{-1/2})$.

  This upper-bound is also tight for certain function classes.

> **Example.** For 1-bounded and 1-Lipschitz functions on $[0, 1]^d$, the uniform $L_2$ entropy (in fact the $L_\infty$ entropy) grows like $(1/\epsilon)^d$.
>
> **Proof.** Since $|f(x) - f(y)| \leq \|x - y\|_2$, $O(n^{-1/d})$ rate here is tight for $d > 2$. ⊛

- $p = 2$: From the same calculation, we have
$$\mathbb{E}\left[\sup_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)\right] \leq \delta + \frac{1}{\sqrt{n}}\int_\delta^1 \frac{C}{\epsilon}\,\mathrm{d}\epsilon = \delta + \frac{C}{\sqrt{n}}\ln\frac{1}{\delta} = O\left(\frac{1}{\sqrt{n}}\log n\right)$$
  by setting $\delta = O(1/\sqrt{n})$. Hence, the Rademacher complexity is bounded by $\log n/\sqrt{n}$, which is "almost" the parametric rate up to the extra log factor. This might not be tight.

**Remark.** To summarize, we have the following bounds on the Rademacher complexity:

- Parametric class: $C\sqrt{d/n}$.
- Non-parametric class:
  - $p < 2$: $C\sqrt{1/n}$;
  - $p = 2$: $C \log n/\sqrt{n}$;
  - $p > 2$: $C \cdot n^{-1/p}$.

**Example** (Linear function class). Let $\chi = B_2^d$, and $\mathscr{F} = \{x \mapsto w^\top x \colon w \in B_2^d\}$. For a given data $x_1, \dots, x_n \in \mathbb{R}^d$,

$$\mathscr{F}|_{x_1,\dots,x_n} = \left\{ Xw \colon w \in B_2^d, X_{n\times d} = \begin{bmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{bmatrix} \right\}.$$

To determine whether $\mathscr{F}$ is parametric or non-parametric, we need to bound $N(\mathscr{F}, L_2(\mathbb{P}_n), \epsilon)$:

$$\begin{aligned}
&\sqrt{\frac{1}{n}\sum_{i=1}^n (\langle w, x_i\rangle - \langle w', x_i\rangle)^2} && && N(\mathscr{F}, L_2(\mathbb{P}_n), \epsilon) \\
&\le \max_{i\in[n]} |\langle w - w', x_i\rangle| && && \le N(\mathscr{F}, L_\infty(\mathbb{P}_n), \epsilon) \\
&\le \max_{x\in B_2^d} |\langle w - w', x\rangle| && \Leftrightarrow && \le N(\mathscr{F}, \|\cdot\|_\infty, \epsilon) \\
&\le \|w - w'\|_2 && && \le N(B_2^d, \|\cdot\|_2, \epsilon) \\
&\le \epsilon, &&
\end{aligned}$$

where $\max_{x\in B_2^d} |\langle w - w', x\rangle| \le \|w - w'\|_2$ since $\|x\|_2 \le 1$. Then, from Proposition 3.3.1,

$$N(B_2^d, \|\cdot\|_2, \epsilon) \le \left(1 + \frac{2}{\epsilon}\right)^d,$$

so we get a $\sqrt{d/n}$ rate since this satisfies parametric condition.[a]

[a] In high dimension situation, this bound can be loose.

It turns out that one can use the norm constraints to also show that

$$\sup_\mu \log N(\mathscr{F}, L_2(\mathbb{P}_n), \epsilon) \le \frac{C}{\epsilon^2},$$

i.e., a dimension-free bound, hence $\mathscr{F}$ also behaves like a non-parametric class! This bound will be useful in a high dimensional setting when $d$ is not small compared to $n$. Thus, we make this an important remark.

**Remark.** A function class can be viewed as parametric and non-parametric at the same time.

There are other examples.

**Example.** Neural networks are like this: we can either measure its complexity by the *number of parameters* or get dimension-independent bounds on its Rademacher complexity by using *norm constraints*.

**Problem.** For what function classes can be bound in terms of uniform $L_2$ entropy?

**Answer.** We have seen boolean function classes with finite VC dimension, and Hölder smooth function classes. ⊛

# Lecture 16: Beyond VC Dimension: Fat-Shattering Dimension

## 3.3.9 Fat-Shattering Dimension

4 Oct. 9:00

To generalize VC dimension, consider the following.

**Definition.** Let $\mathscr{F}$ be a real-valued function class on $\chi$.

**Definition 3.3.18** ($\epsilon$-shattered)**.** A set $\{x_1, \dots, x_n\}$ of $\chi$ is *$\epsilon$-shattered* by $\mathscr{F}$ if there exists $t_1, \dots, t_n$ such that for all $S \subseteq [n]$, there exists $f \in \mathscr{F}$ such that

$$\begin{cases} f(x_s) \le t_s, & \text{if } s \in S; \\ f(x_s) \ge t_s + \epsilon, & \text{if } s \notin S. \end{cases}$$

**Definition 3.3.19** (Fat-shattering dimension)**.** The *fat-shattering dimension* $\mathrm{VC}(\mathscr{F}, \epsilon)$ of $\mathscr{F}$ on $\chi$ is the maximum integer $D$ such that there exists a size $D$ finite set $A \subseteq \chi$ $\epsilon$-shattered by $\mathscr{F}$.

**Remark.** $\mathrm{VC}(\mathscr{F}, \epsilon)$ is a non-increasing function of $\epsilon$.

**Proof.** IF a set is $\epsilon$-shattered then for any $\delta \le \epsilon$, it's also $\delta$-shattered. ⊛

**Note.** If $\mathscr{F}$ is boolean, then $\mathrm{VC}(\mathscr{F}, \epsilon) = \mathrm{VC}(\mathscr{F})$ for all $\epsilon \in (0, 1)$ by setting $t_s = 0$ for all $s$.

**Example.** Consider $\mathcal{M} = \{f : I \to [-1, 1] \text{ non-decreasing}\}$. Then for all $\epsilon > 0$, $\mathrm{VC}(\mathcal{M}, \epsilon) \le 1 + 2/\epsilon$.

**Definition 3.3.20** (Total variation)**.** The *total variation* of a function $f : \mathbb{R} \supseteq I \to \mathbb{R}$ on an interval $I$ is defined as

$$\mathrm{TV}(f) \coloneqq \sup_{n \ge 1} \sup_{x_1 < \cdots < x_n \in I} \sum_{i=1}^{n} |f(x_i) - f(x_{i-1})|.$$

**Intuition.** The total variation is some measure of smoothness of functions. It's more general than differentiability since $\mathrm{TV}(f)$ can also be defined for discontinuous functions.

**Remark.** Total variation is actually a norm, i.e., triangle inequality holds.

Let's first see one example.

**Example.** Consider $\mathrm{BV}(2) \coloneqq \{f : \mathrm{TV}(f) \le 2\}$. If $f'(x)$ exists, then $\mathrm{TV}(f) = \int |f'(x)| \, \mathrm{d}x$. We see that $\mathrm{BV}(2) \supseteq \mathcal{M}$ since for any non-decreasing function $f$ ranging in $[a, b]$, $\mathrm{TV}(f) = b - a$.

In general, we have the following.

**Lemma 3.3.4.** Let $\mathscr{F} \ni f : I \to \mathbb{R}$ with $\mathrm{TV}(f) \le v$ for all $f \in \mathscr{F}$, i.e., $\mathscr{F} = \mathrm{BV}(v)$. Then

$$\mathrm{VC}(\mathrm{BV}(v), \epsilon) = 1 + \left\lfloor \frac{v}{\epsilon} \right\rfloor.$$

**Proof.** We prove this by proving two directions.

**Claim.** $\mathrm{VC}(\mathrm{BV}(v), \epsilon) \le 1 + \lfloor v/\epsilon \rfloor$.

**Proof.** Let $\{x_1, \dots, x_n\}$ be $\epsilon$-shattered by $\mathscr{F}$, then there exists $t_1, \dots, t_n$ and $f_1, f_2$ such that

$$\begin{cases} f_1(x_i) \le t_i, & \text{if } i \text{ is odd;} \\ f_1(x_i) \ge t_i + \epsilon, & \text{if } i \text{ is even;} \end{cases}, \quad \begin{cases} f_2(x_i) \le t_i, & \text{if } i \text{ is even;} \\ f_2(x_i) \ge t_i + \epsilon, & \text{if } i \text{ is odd.} \end{cases}$$

Consider $f = (f_1 - f_2)/2$, then

$$\begin{cases} f(x_i) \le -\epsilon/2, & \text{if } i \text{ is odd;} \\ f(x_i) \ge \epsilon/2, & \text{if } i \text{ is even;} \end{cases} \Rightarrow \mathrm{TV}(f) \ge (n-1)\epsilon$$

by considering this particular partition $\{x_i\}$ of $I$. Furthermore, since TV is a norm,

$$\mathrm{TV}(f) \le \frac{\mathrm{TV}(f_1) + \mathrm{TV}(f_2)}{2} \le v$$

from triangle inequality, hence $(n-1)\epsilon \le v$, i.e., $n \le 1 + \lfloor v/\epsilon \rfloor$. ⊛

**Claim.** $\mathrm{VC}(\mathrm{BV}(v), \epsilon) \ge 1 + \lfloor v/\epsilon \rfloor$.

**Proof.** Let $d = \lfloor v/\epsilon \rfloor$, and consider $y_1 < y_2 < \cdots < y_d$, which induces $d+1$ intervals

$$I_0 = (-\infty, y_1), \quad I_j = [y_j, y_{j+1}), \quad I_d = [y_d, \infty)$$

Let $\mathcal{G}$ be the set of piece-wise continuous functions on $I_0, \dots, I_d$ taking values between $\{0, \epsilon\}$, so $|\mathcal{G}| = 2^{d+1}$. Then, the set

$$\{x_1, \dots, x_{d+1} \colon x_j \in I_{j-1}\}$$

is $\epsilon$-shattered by $\mathcal{G}$. Finally, since $\mathrm{TV}(g) \le d\epsilon \le v$ for all $g \in \mathcal{G}$, we're done. ⊛

■

**Example** (Linear function class). Let $\chi = B_2^d$, and $\mathscr{F} = \{x \mapsto w^\top x \colon w \in B_2^d\}$. Then $\mathrm{VC}(\mathscr{F}, \epsilon) \le d$; and if we consider $\mathrm{sgn}(w^\top x)$, we get $\mathrm{VC}(\mathscr{F}, \epsilon) = d + 1$.

Consider the following result (which we will not prove).

**Theorem 3.3.6** (Mendelson-Vershynin). Let $\mathscr{F}$ be a class of functions that is uniformly bounded by 1. Then there exists $c > 1$ such that for every $0 < \epsilon \le 1$,

$$\sup_\mu M(\mathscr{F}, L_2(\mu), \epsilon) \le \left(\frac{2}{\epsilon}\right)^{c\,\mathrm{VC}(\mathscr{F}, \frac{\epsilon}{c})}.$$

**Remark.** For $\mathrm{BV}(2)$, Mendelson-Vershynin thoerem gives $\sup_\mu M(\mathrm{BV}(2), \epsilon, L_2(\mu)) \le \exp\left(\frac{c}{\epsilon} \log 2\epsilon\right)$.[a]

[a] Note that $\log 2\epsilon$ is superfluous again.

# Lecture 17: Perceptron Algorithm

6 Oct. 9:00

**As previously seen.** In the previous example, we state that the fat-shattering dimension for a linear function class is $\mathrm{VC}(\mathcal{F}, \epsilon) \le d$, which is not dimension-free. 1

If we further impose a norm constraint $\|w\|_2$, then a dimension-free bound can be obtained; specifically, for all $\epsilon > 0$, we can show that

$$\mathrm{VC}(\mathcal{F}, \epsilon) \le \frac{C}{\epsilon^2}$$

where $C$ is some constant. To prove the above, we can use the perceptron algorithm.

**Algorithm 3.1:** Perceptron Algorithm

**Data:** A data sequence $\{(x_i, y_i)\}_{i=1}^T$, observed one-by-one
**Result:** A linear function with weight $w$

$\hat{w}_1 \leftarrow 0$
**for** $t = 1, \dots, T$ **do**
  observe $x_t \in \chi$ // data
  $\hat{y}_t \leftarrow \operatorname{sgn}(\hat{w}_t^\top x_t)$ // predict $\hat{y}_t$
  observe $y_t \in \{\pm 1\}$ // true label
  $\hat{w}_{t+1} \leftarrow \hat{w}_t + \mathbb{1}_{\hat{y}_t \neq y_t} y_t x_t$
**return** $\hat{w}_{T+1}$

**Remark.** Suppose there exists $w$ such that $y_t = 1$ whenever $w^\top x_t \geq 0$, and $y_t = -1$ whenever $w^\top x_t < 0$, then $y_t w^\top x_t > 0$.

The following lemma (see, e.g., [Nov62]) provides an error bound for the perceptron algorithm.

**Lemma 3.3.5** (Perceptron Mistake Bound). For any sequence $(x_1, y_1), \dots, (x_T, y_T) \in B_2^d \times \{\pm 1\}$, the perceptron algorithm makes at most $1/\gamma^2$ mistakes, where

$$\gamma = \max_{w \in B_2^d} \min_{1 \leq t \leq T} y_t w^\top x_t$$

is the margin of the data sequence $\{(x_i, y_i)\}_{i=1}^T$.

**Proof.** Let $M$ be the total number of mistakes made when running the perceptron algorithm. Suppose a mistake is made in round $t$, then

$$\|\hat{w}_{t+1}\|^2 = \|\hat{w}_t + y_t x_t\|^2 = \|\hat{w}_t\|^2 + 2\underbrace{\langle \hat{w}_t, y_t x_t \rangle}_{\leq 0} + \|x_t\|^2 \leq \|\hat{w}_t\|^2 + 1,$$

implying that $\|\hat{w}_{t+1}\| \leq \sqrt{M}$.

Now, consider the margin $\gamma$: when $\gamma$ is achieved at $w = w^*$, $\gamma \leq y_t (w^*)^\top x_t$; moreover, if there is a mistake at round $t$,

$$\gamma \leq y_t (w^*)^\top x_t = (w^*)^\top (\hat{w}_{t+1} - \hat{w}_t)$$

since $y_t x_t = \hat{w}_{t+1} - \hat{w}_t$ in this case. Summing the above over $t$ results in a telescoping sum

$$M\gamma \leq (w^*)^\top \hat{w}_{T+1} \leq \|w^*\| \|\hat{w}_{T+1}\| \leq \|\hat{w}_{T+1}\| \leq \sqrt{M},$$

hence $M \leq 1/\gamma^2$. ∎

We are now ready to prove the following.

**Theorem 3.3.7.** Let $\chi = B_2^d$, and $\mathscr{F} = \{x \mapsto w^\top x : w \in B_2^d\}$. Then for all $\epsilon > 0$,

$$\mathrm{VC}(\mathscr{F}, \epsilon) \leq \frac{4}{\epsilon^2}.$$

**Proof.** Suppose $\{x_1, \dots, x_T\}$ is $\epsilon$-shattered by $\mathcal{F}$. Then, there exists $t_1, \dots, t_n \in [-1, 1]$ such that for all $S \subset \{1, \dots, n\}$, there exists $w_S \in B_2^d$ such that

$$\begin{cases} w_S^\top x_i \geq t_i + \dfrac{\epsilon}{2}, & \text{if } i \in S; \\ w_S^\top x_i \leq t_i - \dfrac{\epsilon}{2}, & \text{if } i \notin S. \end{cases}$$

Let $\widetilde{x}_i = (x_i, t_i) \in \mathbb{R}^{d+1}$ and $\widetilde{w}_S = (w_S, -1) \in \mathbb{R}^{d+1}$, we can rewrite the above as

$$\begin{cases} \widetilde{w}_S^\top \widetilde{x}_i \geq \dfrac{\epsilon}{2}, & \text{if } i \in S; \\ \widetilde{w}_S^\top \widetilde{x}_i \leq -\dfrac{\epsilon}{2}, & \text{if } i \notin S. \end{cases}$$

Equivalently, for any sign vector $y \in \{\pm 1\}^T$, there exists $\widetilde{w}_y$ such that for all $i$, $y_i \widetilde{w}_y^\top \widetilde{x}_i \geq \epsilon/2$, i.e., the margin $\gamma \geq \epsilon/2$. This means that if we run the perceptron algorithm with $\{\widetilde{x}_i, y_i\}_{i=1}^T$ such that

$$y_i = -\hat{y}_i = \operatorname{sgn}(\hat{w}_t^\top x_t),$$

i.e., the perceptron algorithm makes mistake every round, i.e., $M = T$. But since $\gamma \geq \epsilon/2$, $M \leq 4/\epsilon^2$ from the perceptron mistake bound. Combining these two, we have $T = M \leq 4/\epsilon^2$ if a size $T$ subset of $\chi$ is $\epsilon$-shattered by $\mathcal{F}$, i.e., $\mathrm{VC}(\mathscr{F}, \epsilon) \leq 4/\epsilon^2$. ■

**Remark.** We can try to use the above together with the Mendelson-Vershynin theorem.

**Remark.** Suppose $f_1, \ldots, f_d$ are linearly independent. If $\mathcal{F} = \{\sum_{i=1}^d c_i f_i\}$, $\mathrm{VC}(\mathcal{F}, \epsilon) = d$ for all $\epsilon > 0$.

# Lecture 18: Beyond Uniform Entropy Bound: Bracketing Bound

## 3.4 Bracketing Bound

11 Oct. 9:00

**As previously seen.** So far, we have the uniform entropy bound

$$\mathbb{E}\left[\sup_{f \in \mathscr{F}} |\mathbb{P}_n f - \mathbb{P} f|\right] \leq \frac{c}{\sqrt{n}} \|F\|_{L_2(\mathbb{P})} \int_0^1 \sqrt{\log \sup_\mu N(\mathscr{F}, L_2(\mu), \epsilon \|F\|_{L_2(\mathbb{P})})}\, \mathrm{d}x,$$

where $F$ is an envelope of $\mathscr{F}$.

In this lecture, we will see another bound using bracketing (recall Proposition 3.1.1).

### 3.4.1 Bracketing Number

Consider the following.

**Definition 3.4.1** ($\epsilon$-bracket). Given a probability measure $\mathbb{P}$ on $\chi$ and two functions $\ell, u \colon \chi \to \mathbb{R}$, an *$\epsilon$-bracket*, denoted as $[\ell, u]$, is defined as

$$[\ell, u] \coloneqq \{f \colon \chi \to \mathbb{R} \colon \ell(x) \leq f(x) \leq u(x) \text{ for all } x \in \chi\}$$

such that $\|u - \ell\|_{L_2(\mathbb{P})} \leq \epsilon$.[a]

[a] This is the $L_2(\mathbb{P})$ size of $[\ell, u]$, i.e., $\left(\int_\chi (u(x) - \ell(x))^2 \mathbb{P}(\mathrm{d}x)\right)^{1/2}$.

**Definition 3.4.2** (Bracketing number). For every $\epsilon > 0$, the *$\epsilon$-bracketing number* $N_{[\,]}(\mathscr{F}, L_2(\mathbb{P}), \epsilon)$ of a function class $\mathscr{F}$ from $\chi$ to $\mathbb{R}$ is defined as the smallest number of $\epsilon$-brackets such that every $f \in \mathscr{F}$ belongs to only one of the brackets.

**Lemma 3.4.1.** For every $\epsilon > 0$, $N(\mathscr{F}, L_2(\mathbb{P}), \epsilon/2) \leq N_{[\,]}(\mathscr{F}, L_2(\mathbb{P}), \epsilon)$.

**Proof.** Consider $\epsilon$-brackets $[\ell_i, u_i]$ for $i = 1, \ldots, N_{[\,]}(\mathscr{F}, L_2(\mathbb{P}), \epsilon)$, then $\{(\ell_i + u_i)/2\}_i$ forms an

$\epsilon/2$-net since for any $f \in \mathscr{F}$ and any $x \in \chi$,

$$\left\| f - \frac{u_i + \ell_i}{2} \right\|_{L_2(\mathbb{P})} \le \left\| \frac{u_i - \ell_i}{2} \right\|_{L_2(\mathbb{P})} \le \frac{\epsilon}{2}$$

from the fact that $u_i \ge f \ge \ell_i$ and $\|u_i - \ell_i\|_{L_2(\mathbb{P})} \le \epsilon$. ■

Let's see one simple example of computing bracketing functions.

**Example.** Let $\mathscr{F} = \{\mathbb{1}_{[-\infty,t]} : t \in \mathbb{R}\}$ and $\mathbb{P}$ be a probability measure on $\mathbb{R}$. Then for all $\epsilon > 0$,

$$N_{[\,]}(\mathscr{F}, L_2(\mathbb{P}), \epsilon) \le 1 + 1/\epsilon^2.$$

**Proof.** We show this by finding a collection of $\sqrt{\epsilon}$-brackets with at most $1+1/\epsilon$ many of the brackets. Let $t_0 = -\infty$, and recursively define

$$t_i = \sup\{x : x > t_{i-1} : \mathbb{P}((t_{i-1}, x]) \le \epsilon\}.$$

Finally, let $k \ge 1$ be the smaller integer such that $t_k = \infty$. We then have

- $\mathbb{P}((t_{i-1}, t_i)) \le \epsilon$: for every $\delta > 0$, $\mathbb{P}((t_{i-1}, t_i - \delta]) \le \epsilon$, as $\delta \to 0$, $\mathbb{P}((t_{i-1}, t_i)) \le \epsilon$.
- $\mathbb{P}((t_{i-1}, t_i]) \ge \epsilon$: for every $\delta > 0$, $\mathbb{P}((t_{i-1}, t_i + \delta]) > \epsilon$, as $\delta \to 0$, $\mathbb{P}((t_{i-1}, t_i]) \ge \epsilon$.

Then,

$$1 = \mathbb{P}((-\infty, \infty)) \ge \sum_{i=1}^{k} \mathbb{P}((t_{i-1}, t_i]) \ge (k-1)\epsilon,$$

implying $k \le 1 + \frac{1}{\epsilon}$. Now, consider brackets $[\mathbb{1}_{(-\infty,t_{i-1})}, \mathbb{1}_{(-\infty,t_i)}]$ which cover $\mathscr{F}$ with $L_2(\mathbb{P})$ size equal to $\sqrt{\mathbb{P}((t_{i-1}, t_i))} \le \sqrt{\epsilon}$. Hence, this is a collection of valid $\sqrt{\epsilon}$-brackets of size $\le 1 + 1/\epsilon$, i.e., by replacing $\epsilon \leftarrow \sqrt{\epsilon}$, we have $N_{[\,]}(\mathscr{F}, L_2(\mathbb{P}), \epsilon) \le 1 + 1/\epsilon^2$. ⊛

**Proposition 3.4.1.** Let $\mathscr{F}$ to be a function class such that $N_{[\,]}(\mathscr{F}, L_2(\mathbb{P}), \epsilon) < \infty$ for all $\epsilon > 0$. Then as $n \to \infty$,

$$\sup_{f \in \mathscr{F}} |\mathbb{P}_n f - \mathbb{P} f| \xrightarrow{\text{a.s.}} 0.$$

**Proof.** Fix $\epsilon > 0$, let $[\ell_i, u_i]$ for $i = 1, \dots, N$ to be a set of $\epsilon$-brackets. Then, it suffices to show[a]

$$\sup_{f \in \mathscr{F}} |\mathbb{P}_n f - \mathbb{P} f| \le \left( \max_{1 \le i \le N} \max(|\mathbb{P}_n u_i - \mathbb{P} u_i|, |\mathbb{P}_n \ell_i - \mathbb{P}\ell_i|) \right) + \epsilon.$$

To show this, let $f \in [\ell_i, u_i]$ for some $i$, then

$$\mathbb{P}_n f - \mathbb{P} f \le (\mathbb{P}_n u_i - \mathbb{P} u_i) + (\mathbb{P} u_i - \mathbb{P} f) \le (\mathbb{P}_n u_i - \mathbb{P} u_i) + \mathbb{P}(u_i - \ell_i) \le \mathbb{P}_n u_i - \mathbb{P} u_i + \epsilon$$

since $\mathbb{P}(u_i - \ell_i) \le \|u_i - \ell_i\|_{L_2(\mathbb{P})} \le \epsilon$. On the other hand, we also have

$$\mathbb{P} f - \mathbb{P}_n f \le (\mathbb{P} f - \mathbb{P}\ell_i) + (\mathbb{P}\ell_i - \mathbb{P}_n \ell_i) \le (\mathbb{P} u_i - \mathbb{P}\ell_i) + (\mathbb{P}\ell_i - \mathbb{P}_n \ell_i) \le |\mathbb{P}_n \ell_i - \mathbb{P}\ell_i| + \epsilon,$$

hence we're done. ■

[a] It then implies $\limsup_{n\to\infty} |\mathbb{P}_n f - \mathbb{P} f| \le \epsilon$ almost surely just by the law of large number. By taking $\epsilon = 1/m$ to 0, we can say that $|\mathbb{P}_n f - \mathbb{P} f| \to 0$ almost surely.

### 3.4.2 Bracketing Bound

The main theorem of this section is the following.

**Theorem 3.4.1** (Bracketing bound). Let $F$ be an envelope of $\mathscr{F}$ such that $\mathbb{P}F^2 < \infty$. Then for some constant $C > 0$,

$$\mathbb{E}\left[\sup_{f\in\mathscr{F}} \sqrt{n}(\mathbb{P}_n f - \mathbb{P}f)\right] \le C\|F\|_{L_2(\mathbb{P})} \int_0^1 \sqrt{\log N_{[\,]}(\mathscr{F}, L_2(\mathbb{P}), \epsilon\|F\|_{L_2(\mathbb{P})})}\,\mathrm{d}\epsilon.$$

**Remark.** The main differences between the bracketing bound and the uniform entropy bound are

- covering number is replaced by bracketing number;
- We do not have the $\sup_\mu$, hence the bracketing bound is only w.r.t. $\mathbb{P}$.

# Lecture 19: Applications to $M$-Estimators

13 Oct. 9:00

The following shows that using the bracketing number is more tractable than using the uniform $L_2$ covering number.

**Lemma 3.4.2** (Parametric Lipschitz functin class). Let $\Theta \subseteq \mathbb{R}^d$ be constrained in an $L_2$ ball of radius $R$, and let $\mathscr{F} = \{m_\theta \colon \chi \to \mathbb{R} \colon \theta \in \Theta\}$ be a function class indexed by $\theta$. Suppose there exists a function $M(x)$ with $\|M\|_{L_2(\mathbb{P})} < \infty$ such that

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \le M(x)\|\theta_1 - \theta_2\|_2$$

for all $x \in \chi$ and $\theta_1, \theta_2 \in \Theta$. Then, for all $\epsilon > 0$,

$$N_{[\,]}(\mathscr{F}, L_2(\mathbb{P}), \epsilon\|M\|_{L_2(\mathbb{P})}) \le \left(1 + \frac{4R}{\epsilon}\right)^d.$$

**Proof.** Let $\{\theta_i\}_{i=1}^N$ be a maximal $\epsilon/2$-packing of $\Theta$, and consider the following brackets:

**Claim.** The brackets $[m_{\theta_i} \pm \epsilon M/2]$ for $i = 1, \dots, N$ cover $\mathscr{F}$.

**Proof.** First, note that a maximal packing set is indeed a covering net with the same $\epsilon$. Therefore, for all $m_\theta \in \mathscr{F}$, there exists $i$ such that $\|\theta - \theta_i\|_2 \le \epsilon/2$, hence

$$|m_\theta(x) - m_{\theta_i}(x)| \le M(x)\|\theta - \theta_i\|_2 \le M(x)\frac{\epsilon}{2},$$

for all $x \in \chi$, i.e., $m_\theta \in [m_{\theta_i} \pm \epsilon M/2]$. This means the brackets $[m_{\theta_i} \pm \epsilon M/2]$ cover $\mathscr{F}$. ⊛

Furthermore, the size of each bracket is $\epsilon\|M\|_{L_2(\mathbb{P})}$, with Proposition 3.3.1, we're done. ■

Some examples of the parametric Lipschitz function classes are the following.

**Example.** For $m_\theta(x) = \theta^\top x$, $|\theta_1^\top x - \theta_2^\top x| \le \|\theta_1 - \theta_2\|_2\|x\|_2$ from Cauchy-Schwarz. Hence, $M(x) = \|x\|_2$. For Lemma 3.4.2 to apply, consider $\mathbb{P}$ such that $\mathbb{P}\|x\| < \infty$.

**Example** (Quantile regression). For $m_\theta(x) = |x - \theta|$, $|m_{\theta_1}(x) - m_{\theta_2}(x)| \le |\theta_1 - \theta_2|$ with $M(x) = 1$. In this case, since any measure $\mathbb{P}$ gives $\mathbb{P}1 < \infty$, so Lemma 3.4.2 applies for all $\mathbb{P}$.

The above examples extend to essentially all $p$-norm.

**Example.** $m_\theta(x) = \|x - \theta\|_p$.

Finally, let's see some standard results on the bracketing numbers.

**Example** ($\alpha$-Hölder smooth function class)**.** For $\mathcal{S}_\alpha$ on $[0,1] \to [0,1]$,

$$\log N_{[\,]}(\mathcal{S}_\alpha, L_2(\mathbb{P}), \epsilon) \le C \left(\frac{1}{\epsilon}\right)^\alpha.$$

**Example.** Let $\mathcal{M}$ be the monotone function class on $\mathbb{R} \to [0,1]$,

$$\log N_{[\,]}(\mathcal{M}, L_2(\mathbb{P}), \epsilon) \le C \left(\frac{1}{\epsilon}\right).$$

**Example.** Consider $\mathcal{C}$ be the set of convex function class on $[0,1] \to [0,1]$. Let $\mathcal{U}$ be the uniform distribution on $[0,1]$. Then

$$\log N_{[\,]}(\mathscr{C}, L_2(\mathcal{U}), \epsilon) \le C \left(\frac{1}{\sqrt{\epsilon}}\right).$$

More examples are available in [VW96]. To conclude this section, we ask the following:

**Problem 3.4.1** (Necessity of VC dimension of boolean function class)**.** Is there a function class for which the uniform entropy bound is infinite, while the bracketing bound is finite?

**Answer.** Yes! For boolean function class $\mathscr{F}$, from the Dudley's theorem,

- $\mathrm{VC}(\mathscr{F}) < \infty$: $\sup_\mu N(\mathscr{F}, L_2(\mu), \epsilon) < \infty$;
- $\mathrm{VC}(\mathscr{F}) = \infty$: $\sup_\mu N(\mathscr{F}, L_2(\mu), \epsilon) = \infty$ for all $\epsilon < 1/2$.

In addition, if $\mathrm{VC}(\mathscr{F}) = \infty$, uniform Glivenko-Cantelli does not hold, i.e.,

$$\liminf_{n\to\infty} \sup_{\mathbb{P}} \mathbb{E}\left[\sup_{f\in\mathscr{F}} \mathbb{P}_n f - \mathbb{P} f\right] > 0.$$

However, it's still possible that when $\mathrm{VC}(\mathscr{F}) = \infty$, $\mathscr{F}$ is Glivenko-Cantelli w.r.t. some $\mathbb{P}$. ⊛

**Example.** Let $\mathscr{F} = \{\mathbb{1}_C : C \text{ is compact convex subset of } [0,1]^d\}$. Then $\mathrm{VC}(\mathscr{F}) = \infty$, and $\mathscr{F}$ is Glivenko-Cantelli w.r.t. any $\mathbb{P}$ having a density w.r.t. the Lebesgue measure.

**Lemma 3.4.3.** Let $\mathscr{F}$ be a class of functions uniformly bounded by 1. Then for every $\epsilon > 0$,

$$\frac{1}{8}\mathrm{VC}(\mathscr{F}, 4\epsilon) \le \log \sup_\mu N(\mathscr{F}, L_2(\mu), \epsilon)$$

# Chapter 4

# Applications to $M$-Estimation

In this chapter, we will focus on $M$-estimator, and primarily investigate the "rate of convergence" for $M$-estimators, and look at some examples.

## 4.1 The $M$-Estimation Problem

Consider the problem of $M$-estimation, which formalize subsection 1.2.1:

**Problem 4.1.1** ($M$-estimation)**.** Let $\Theta$ be an abstract parameter space,[a] and let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$ be the data. Let $M\colon \Theta \to \mathbb{R}$ and $M_n\colon \Theta \to \mathbb{R}$ be random functions[b] depend on the data. Then, the *$M$-estimation* problem tries to estimate the true parameter

$$\theta_0 = \arg\max_{\theta\in\Theta} M(\theta)$$

by minimizing $M_n(\theta)$ instead and find

$$\hat{\theta}_n = \arg\max_{\theta\in\Theta} M_n(\theta).$$

[a] E.g., $\mathbb{R}^d$, or some function spaces.
[b] Or equivalently, one can view $\{M_n\}_n$ as a stochastic process.

**Remark.** Typically, for each fixed $\theta \in \Theta$, we have $M_n(\theta) \xrightarrow{p} M(\theta)$.

We have seen some examples in the beginning of the class (subsection 1.2.1).

**Example** (§1.2.1)**.** Mean, Quantile, and Mode estimation.

**Example** (Least square estimation)**.** Let $(X_1, Y_1), \dots, (X_n, Y_n) \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, and let $\Theta = \mathscr{F}$. Consider $M(f) = -\mathbb{P}(y - f(x))^2$ and $M_n(f) = -\mathbb{P}_n(y - f(x))^2$.

**Example** (ERM in classification)**.** Let $M(f) = -\mathbb{P}(y \neq \operatorname{sgn}(f(x)))$ and $M_n(f) = -\mathbb{P}_n(y \neq \operatorname{sgn}(f(x)))$.

**Example** (MLE)**.** Let $M(\theta) = p \log p_\theta$ and $M_n(\theta) = p_n \log p_\theta$, where $\{p_\theta\}_{\theta\in\Theta}$ is a class of densities w.r.t. some measure.

It only makes sense to look at those $M$-esmator that are consistent.

**Definition 4.1.1** (Consistent)**.** An $M$-esmator $\hat{\theta}$ is *consistent* if $|M(\hat{\theta}_n) - M(\theta_0)| \xrightarrow{p} 0$ implies $d(\hat{\theta}_n, \theta_0) \xrightarrow{p} 0$ as $n \to \infty$.

Then, we can ask the following three questions (progressively harder) for an $M$-estimator.

**Problem 4.1.2** (Consistency). Is the $M$-estimator consistent? I.e., as $|M(\hat{\theta}_n) - M(\theta_0)| \xrightarrow{p} 0$, does $d(\hat{\theta}_n, \theta_0) \xrightarrow{p} 0$ for some metric $d$?

**Problem 4.1.3** (Rate of convergence). What's the "rate of convergence" for the $M$-estimator?

**Example.** The rate for the mean is $O(1/\sqrt{n})$.

We will define the "rate of convergence" precisely later (Definition 4.3.1).

**Problem 4.1.4** (Limiting distribution). What's the limiting (asymptotic) distribution of $\hat{\theta} - \theta_0$?

### 4.1.1 Running Example

In the remaining class, we will consider the sample mode estimation as our running example.

**Example** (Mode estimation). Let $\chi = \Theta = \mathbb{R}$, and suppose we have $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{\theta_0}$ supported on $\chi$ such that $\mathbb{P}_{\theta_0}$ has smooth and bounded density $p_{\theta_0}(x)$ w.r.t. Lebesgue measure, with $p'_{\theta_0}(x) > 0$ for $x < \theta_0$, and $p'_{\theta_0}(x) < 0$ for $x > \theta_0$. The *mode estimation* problem considers

$$M(\theta) = \mathbb{P}_{\theta_0}(\theta - 1 \le X \le \theta + 1),$$

so the true parameter $\theta_0 = \arg\max_{\theta \in \Theta} M(\theta)$ is the mode. Let

$$M_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\theta - 1 \le X_i \le \theta + 1},$$

and let $\hat{\theta}_n = \arg\max_{\theta \in \Theta} M_n(\theta)$, i.e., the sample mode.

**Remark** (Unique optimal). Notice that in the mode estimation problem, we can check that

- $M'(\theta_0) = 0$;
- $M'(\theta) < 0$ when $\theta < 0$, and $M'(\theta) > 0$ when $\theta > 0$;
- $M''(\theta) > 0$.

These conditions guarantee that $\theta_0$ is the unique maximum.

## 4.2 Consistency

Consistency is the easiest task among others. Firstly, we need to show that $|M(\hat{\theta}_n) - M(\theta_0)| \xrightarrow{p} 0$: recall the "basic inequality" step, i.e.,

$$|M(\hat{\theta}_n) - M(\theta_0)| = |M(\hat{\theta}_n) - M_n(\hat{\theta}_n) + \underbrace{M_n(\hat{\theta}_n) - M_n(\theta_0)}_{\le 0} + M_n(\theta_0) - M(\theta_0)| \le 2 \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)|.$$

In the mode estimation example, denote $m_\theta(x) = \mathbb{1}_{\theta - 1 \le x \le \theta + 1}$, and $\mathscr{F} = \{m_\theta : \theta \in \mathbb{R}\}$, then consider the following notation.

**Notation.** $M_n(\theta) = \mathbb{P}_n m_\theta$ and $M(\theta) = \mathbb{P} m_\theta$.

We then have

$$\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| = \sup_{\theta \in \mathbb{R}} |\mathbb{P}_n m_\theta - \mathbb{P} m_\theta| \xrightarrow{p} 0$$

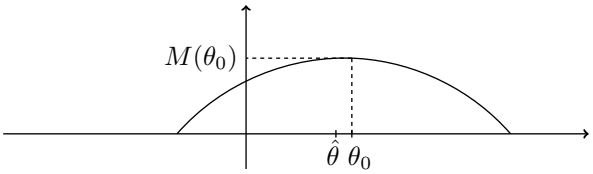
since $\mathrm{VC}(\mathscr{F}) = 2$ (see the previous example). Hence, we conclude that $M(\hat{\theta}) - M(\theta_0) \xrightarrow{p} 0$ as $n \to \infty$. Now, it remains to answer the following.

**Problem.** Does $\hat{\theta} \xrightarrow{p} \theta_0$, i.e., does $d(\hat{\theta}, \theta_0) \xrightarrow{p} 0$?

**Answer.** Need to relate $d$ to $M$ function. ⊛

# Lecture 20: A Heuristic Argument for Rate of Convergence

To show $\hat{\theta} \xrightarrow{p} \theta_0$, we need "curvature" or "growth" condition on $M$ at $\theta_0$. 16 Oct. 9:00




Consider the following.

**Definition 4.2.1** (Growth condition). The *growth condition* on $M(\theta)$ is that for all $\epsilon > 0$,

$$\sup_{\theta\colon\, d(\theta,\theta_0)>\epsilon} M(\theta) < M(\theta_0).$$

Equivalently, the growth condition implies that for all $\epsilon > 0$, there exists $\delta > 0$ such that $\delta < \epsilon$ and

$$\inf_{\theta\colon\, d(\theta,\theta_0)>\epsilon} M(\theta_0) - M(\theta) > \delta.$$

Hence, as $\epsilon \to 0$, $\mathbb{P}(d(\hat{\theta}, \theta_0) > \epsilon) \le \mathbb{P}(M(\theta_0) - M(\hat{\theta}) > \delta) \to 0$. In our mode estimation example, since

$$M(\theta) = \mathbb{P}(\theta - 1 \le X \le \theta + 1) = \int_{\theta-1}^{\theta+1} p_{\theta_0}(x)\,\mathrm{d}x,$$

if we assume that $p_{\theta_0}$ is increasing until $\theta_0$ and decreasing after $\theta_0$,[1] one can check that

$$\sup_{\theta\colon\, d(\theta,\theta_0)>\epsilon} M(\theta) = \max(M(\theta_0 + \epsilon), M(\theta_0 - \epsilon)) < M(\theta_0)$$

from $p'' < 0$ at $\theta_0$. More generally, one can refer to the following [Vaa98, Theorem 5.7].

**Theorem 4.2.1.** Let $\hat{\theta}$ be any minimizer of $M_n(\theta)$, and $\theta_0$ be the unique minimizer of $M(\theta)$. If

(a) $\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{p} 0$;

(b) for all $\epsilon > 0$, $\inf_{\theta\colon\, d(\theta,\theta_0)>\epsilon} M(\theta) > M(\theta_0)$ for some metric $d$ on $\Theta$,

then we have $d(\theta, \theta_0) \xrightarrow{p} 0$.

**Remark.** Condition (a) may be too strong. One may need a preliminary "localization".

## 4.3 Rate of Convergence

On the other hand, if we can directly show a rate of convergence, then we don't even need to show consistency. Let's first see some heuristics before introducing the general theory.

[1] Since $\theta_0$ is the unique maximum.

### 4.3.1 The Heuristic Argument

In this section, we'll first show a heuristic argument for bounding the rate of convergence in the case of our running example, i.e., mode estimation, where $d(\theta, \theta') = |\theta - \theta'|$.

> **As previously seen.** Recall the previously defined notation for the mode estimation example, i.e., $M_n(\theta) = \mathbb{P}_n m_\theta$ and $M(\theta) = \mathbb{P} m_\theta$ where $m_\theta(x) = \mathbb{1}_{\theta-1 \le x \le \theta+1}$.

We then see that in the mode estimation example,

$$M(\theta_0) - M(\hat{\theta}_n) \le (\mathbb{P}_n - \mathbb{P})(m_{\hat{\theta}_n} - m_{\theta_0}) \le 2 \sup_{\theta \in \Theta} |\mathbb{P}_n m_\theta - \mathbb{P} m_\theta|,$$

and we can further upper-bound it by $c/\sqrt{n}$ in expectation, i.e., in expectation,

$$M(\theta_0) - M(\hat{\theta}_n) = O_p(1/\sqrt{n}).$$

Now, our goal is to somehow relate $|\hat{\theta}_n - \theta_0|$ with $M(\theta_0) - M(\hat{\theta}_n)$.

> **Note.** We cannot in general, get a rate better than $1/\sqrt{n}$ for this suprema.

> **Remark.** If $M$ is twice differentiable at $\theta_0$ with $M''(\theta_0) < 0$, then there exists a neighborhood $I$ of $\theta_0$ such that for all $\theta \in I$,
>
> $$M(\theta_0) - M(\theta) \ge c \cdot |\theta - \theta_0|^2.$$
>
> **Proof.** From Taylor expansion, there exists $\xi$ between $\theta$ and $\theta_0$ such that
>
> $$M(\theta) - M(\theta_0) = |\theta - \theta_0| M'(\theta_0) + \frac{|\theta - \theta_0|^2}{2} M''(\xi),$$
>
> with $M'(\theta_0) = 0$, $M(\theta) - M(\theta_0) > c \cdot |\theta - \theta_0|^2$ for some $c$. ⊛

Since $\hat{\theta}_n$ is consistent, we can assume $\hat{\theta}_n \in I$. The upshot is that we want some sorts of "growth condition" (different from Definition 4.2.1)[2] not for all $\theta \in \mathbb{R}$, but for $\theta$ in a fixed neighborhood $I$ of $\theta_0$.

> **Intuition.** Since $\hat{\theta}_n$ is consistent, the growth condition on a ball around $\theta_0$ should suffice.

In this case, the growth condition relating $|\hat{\theta}_n - \theta_0|$ and $M(\theta_0) - M(\hat{\theta}_n)$ is

$$|\hat{\theta}_n - \theta_0|^2 \lesssim M(\theta_0) - M(\hat{\theta}_n).$$

Then by actually showing $M(\theta_0) - M(\hat{\theta}_n) = O_p(1/\sqrt{n})$, we can conclude $|\hat{\theta}_n - \theta_0| = O_p(n^{-1/4})$.

> **Note.** In our example of mode estimation, $M''$ is constant because $p'_{\theta_0}$ is constant, and
>
> $$M''(\theta_0) = p'_{\theta_0}(\theta_0 + 1) - p'_{\theta_0}(\theta_0 - 1) < 0$$
>
> since $p'(\theta_0 + 1) < 0$ and $p'(\theta_0 - 1) > 0$.

Specifically, we now have

$$|\hat{\theta}_n - \theta_0|^2 \lesssim M(\theta_0) - M(\hat{\theta}_n) \le (\mathbb{P}_n - \mathbb{P})(m_{\hat{\theta}_n} - m_{\theta_0}) \le 2 \sup_{\theta \in \Theta} |\mathbb{P}_n m_\theta - \mathbb{P} m_\theta| \le \frac{c}{\sqrt{n}} = O_p\left(\frac{1}{\sqrt{n}}\right)$$

in expectation.

> **Remark** (The right rate). We cannot get better than $O_p(n^{-1/4})$ with this argument. But there exist problems where the rate is better (need to do "localization")!
>
> - In our example of mode estimation, the right rate is $O_p(n^{-1/3})$.

[2] This is precisely defined in Definition 4.3.2.

- For "parametric" problems (e.g., mean/quantile estimation) the right rate is $O_p(1/\sqrt{n})$.

Now comes the heuristic part. Let's first compute the bound for $\mathbb{E}\left[(\mathbb{P}_n - \mathbb{P})(m_\theta - m_{\theta_0})\right]$ for a fixed $\theta$ close to $\theta_0$. We have

$$\begin{aligned}\mathbb{E}\left[|(\mathbb{P}_n - \mathbb{P})(m_\theta - m_{\theta_0})|\right] &\le \sqrt{\operatorname{Var}\left[\mathbb{P}_n(m_\theta - m_{\theta_0})\right]} \\ &= \frac{1}{\sqrt{n}}\sqrt{\operatorname{Var}\left[m_\theta(X_1) - m_{\theta_0}(X_1)\right]} \le \frac{1}{\sqrt{n}}\sqrt{\mathbb{E}\left[\left(m_\theta(X_1) - m_{\theta_0}(X_1)\right)^2\right]}.\end{aligned}$$

In our mode estimation problem, if $\theta$ is close to $\theta_0$ (with $\theta < \theta_0$),

$$m_\theta(x) - m_{\theta_0}(x) = \mathbb{1}_{\theta-1\le x\le\theta_0+1} + \mathbb{1}_{\theta+1\le x\le\theta_0+1},$$

hence

$$\mathbb{E}\left[\left(m_\theta(X_1) - m_{\theta_0}(X_1)\right)^2\right] \le \mathbb{P}(\theta - 1 \le X_1 \le \theta_0 - 1) + \mathbb{P}(\theta + 1 \le X_1 \le \theta_0 + 1) \le 2p_{\theta_0}(\theta_0)\cdot|\theta - \theta_0|,$$

and since $p_{\theta_0}$ is bounded, we finally have

$$\mathbb{E}\left[\left(m_\theta(X_1) - m_{\theta_0}(X_1)\right)^2\right] \lesssim |\theta - \theta_0|,$$

which implies

$$(\mathbb{P}_n - \mathbb{P})(m_\theta - m_{\theta_0}) = O_p\left(\sqrt{\frac{|\theta - \theta_0|}{n}}\right).$$

**Intuition** (Heuristic). Heuristically, we might want to conclude that

$$(\mathbb{P}_n - \mathbb{P})(m_{\hat\theta_n} - m_{\theta_0}) = O_p\left(\sqrt{\frac{|\hat\theta_n - \theta_0|}{n}}\right),$$

which is a better bound than before.

If this is true, then the overall bound becomes

$$|\hat\theta_n - \theta_0|^2 \le O_p\left(\sqrt{\frac{|\hat\theta_n - \theta_0|}{n}}\right).$$

Canceling $\sqrt{|\hat\theta_n - \theta_0|}$ from both sides,

$$|\hat\theta_n - \theta_0|^{3/2} = O_p\left(\frac{1}{\sqrt{n}}\right) \Rightarrow |\hat\theta_n - \theta_0| = O_p(n^{-1/3}),$$

which is the correct rate for $\hat\theta_n - \theta_0$.

**Remark.** In fact, the limiting distribution is also known where we have

$$n^{1/3}(\hat\theta_n - \theta_0) \xrightarrow{d} \arg\max_{h\in\mathbb{R}} aB_h - bh^2,$$

where $a, b$ are constants depend on $p$.

# Lecture 21: The General Argument for Rate of Convergence

## 4.3.2 A General Approach

18 Oct. 9:00

We're now going to show the general argument for bounding the rate of convergence. In this section, we will assume that our $M$-estimation problem is defined for minimum rather than maximum.[3]

[3] This can be done without loss of generality since we can simply add a negative sign for $M$ and $M_n$.

**Note.** Hence, $M(\theta) \geq M(\theta_0)$ for all $\theta \in \Theta$ now.

**Definition 4.3.1** (Rate of convergence)**.** The *rate of convergence* for $\hat{\theta}_n$ is defined as the sequence $\{\delta_n\}$ such that $d(\hat{\theta}_n, \theta_0) = O_p(\delta_n)$.

Recall that instead of using the old growth condition used when showing consistency, we need an alternative form. Consider the following.

**Definition 4.3.2** (Growth condition*)**.** The *growth condition* on $M(\theta)$ is that for all $\theta \in \Theta$,

$$d(\theta, \theta_0)^2 \leq M(\theta) \ - M(\theta_0).$$

**Note.** For such growth condition, the canonical choice of $d$ is $d(\theta, \theta_0) = \sqrt{M(\theta_0) - M(\theta)}$.

It suffices to show that given $\epsilon > 0$, there exists $M$ (not depending on $n$) such that for all $n$,

$$\mathbb{P}(d(\hat{\theta}_n, \theta_0) > 2^M \delta_n) \leq \epsilon.$$

Firstly, we apply the *peeling step* to get

$$\mathbb{P}(d(\hat{\theta}_n, \theta_0) > 2^M \delta_n) = \sum_{j > M} \mathbb{P}(2^{j-1}\delta_n < d(\hat{\theta}_n, \theta_0) \leq 2^j \delta_n). \tag{4.1}$$

**Note.** From the growth condition and the basic inequality,

$$\begin{aligned} d(\hat{\theta}_n, \theta_0)^2 &\leq M(\hat{\theta}_n) - M(\theta_0) \\ &= M(\hat{\theta}_n) - M_n(\hat{\theta}_n) + \underbrace{M_n(\hat{\theta}_n) - M_n(\theta_0)}_{\leq 0} + M_n(\theta_0) - M(\theta_0) \\ &\leq (M_n - M)(\theta_0) - (M_n - M)(\hat{\theta}_n). \end{aligned}$$

With this, we can further upper-bound Equation 4.1 by

$$\begin{aligned} \mathbb{P}(d(\hat{\theta}_n, \theta_0) > 2^M \delta_n) &= \sum_{j > M} \mathbb{P}(2^{j-1}\delta_n < d(\hat{\theta}_n, \theta_0) \leq 2^j \delta_n) && \text{peeling step} \\ &\leq \sum_{j > M} \mathbb{P}((M_n - M)(\theta_0) - (M_n - M)(\hat{\theta}_n) \geq 2^{2j-2}\delta_n^2 \cap d(\hat{\theta}_n, \theta_0) \leq 2^j \delta_n) \\ &\leq \sum_{j > M} \mathbb{P}\left( \sup_{\theta \colon d(\theta, \theta_0) \leq 2^j \delta_n} (M_n - M)(\theta_0) - (M_n - M)(\theta) \geq 2^{2j-2}\delta_n^2 \right) \\ &\leq \sum_{j > M} \frac{\mathbb{E}\left[\sup_{\theta \colon d(\theta, \theta_0) \leq 2^j \delta_n} (M_n - M)(\theta_0) - (M_n - M)(\theta)\right]}{2^{2j-2}\delta_n^2} \end{aligned}$$

by Markov's inequality, where we need to assume that it's non-negative. Now, we define the following.

**Definition 4.3.3** (Localized empirical process)**.** The *localized empirical process* for an $M$-estimaton problem for $t > 0$ is defined as

$$\mathbb{E}\left[ \sup_{\theta \colon d(\theta, \theta_0) \leq t} (M_n - M)(\theta_0) - (M_n - M)(\theta) \right].$$

Note the following.

**Note.** For nearly all $M$-estimaton problems, the localized empirical process can be upper-bounded by a sequence of functions $\phi_n : [0, \infty] \to [0, \infty]$ such that for all $t > 0$,

$$\mathbb{E}\left[\sup_{\theta:\, d(\theta,\theta_0)\le t} (M_n - M)(\theta_0) - (M_n - M)(\theta)\right] \le \phi_n(t).$$

Assuming $\phi_n$'s exist, we then proceed upper-bounding Equation 4.1 as

$$\mathbb{P}(d(\hat{\theta}_n, \theta_0) > 2^M \delta_n) \le \sum_{j>M} \frac{\mathbb{E}\left[\sup_{\theta:\, d(\theta,\theta_0)\le 2^j\delta_n} (M_n - M)(\theta_0) - (M_n - M)(\theta)\right]}{2^{2j-2}\delta_n^2} \le \sum_{j>M} \frac{\phi_n(2^j\delta_n)}{2^{2j-2}\delta_n^2}.$$

To further bound the right-hand side, consider the following.

**Definition 4.3.4** (Sub-quadratic assumption)**.** The *sub-quadratic assumption* assumes that there exists $\alpha < 2$ such that for all $n$, $c > 1$, and $x > 0$,

$$\phi_n(cx) \le c^\alpha \cdot \phi_n(x).$$

With sub-quadratic assumption, we get

$$\mathbb{P}(d(\hat{\theta}_n, \theta_0) > 2^M \delta_n) \le \sum_{j>M} \frac{\phi_n(2^j\delta_n)}{2^{2j-2}\delta_n^2} \le 4 \sum_{j>M} \frac{2^{\alpha j}\phi_n(\delta_n)}{2^{2j}\delta_n^2}.$$

This is the final bound we will get by introducing $\phi_n$. Now, to control them, consider the so-called rate-determining equation.

**Definition 4.3.5** (Rate-determinig equation)**.** Given a sequence $\{\delta_n\}$, the *rate-determining equation* for a localized empirical process's upper-bounds $\phi_n$'s is that for some $c$ such that for all $n$,

$$\phi_n(\delta_n) \le c\delta_n^2.$$

**Remark.** It's important to check that whether such $c$ exists for all $n$.

**Intuition.** We want to have $\phi_n(\delta_n) \approx \delta_n^2$.

Finally, assuming the rate-determining equation exists for some $c$, we get

$$\mathbb{P}(d(\hat{\theta}_n, \theta_0) > 2^M \delta_n) \le 4 \sum_{j>M} \frac{2^{\alpha j}\phi_n(\theta_n)}{2^{2j}\delta_n^2} \le 4c \cdot \sum_{j>M} \frac{2^{\alpha j}}{2^{2j}},$$

and from the sub-quadratic assumption, $\alpha < 2$, so the above sum converges to 0 as $M \to \infty$.

**Remark.** We can choose $M$ not depending on $n$ such that the right-hand side is $\le \epsilon$.

Putting the above together, we have the following.

**Theorem 4.3.1** (Non-asymptotic rate of convergence)**.** For an $M$-estimation problem, assume the growth condition on $M$, and the sub-quadratic assumption (with parameter $\alpha < 2$) and the rate-determining equation are valid for $\phi_n$'s arose from bounding the localized empirical process,

$$\mathbb{P}(d(\hat{\theta}_n, \theta_0) > 2^M \delta_n) \le 4c \sum_{j>M} 2^{(\alpha-2)j}.$$

Back to the mode estimation example, we now want to formally show that the rate of convergence for $|\hat{\theta}_n - \theta_0|$ is $O_p(n^{-1/3})$.

**Proposition 4.3.1.** For the mode estimation problem, the rate of convergence for $\hat{\theta}_n$ is $O_p(n^{-1/3})$.

**Proof.** We check the following.

- The growth condition is checked in a ball around $\theta_0$ and not for all $\theta \in \Theta$ (since it's consistent). This is allowed as shown in Theorem 4.3.2.
- Consider the localized empirical process with notation $m_\theta(x) = \mathbb{1}_{[\theta-1,\theta+1]}$, we have

$$\mathbb{E}\left[\sup_{\theta:\,|\theta-\theta_0|\le t} |\mathbb{P}_n(m_\theta - m_{\theta_0}) - \mathbb{P}(m_\theta - m_{\theta_0})|\right].$$

  Let $f_\theta(x) = m_\theta(x) - m_{\theta_0}(x)$, and $\mathscr{F} = \{f_\theta : |\theta - \theta_0| \le t\}$. One can check that

$$f(x) = \mathbb{1}_{\theta_0-1-t\le x\le \theta_0-1+t} + \mathbb{1}_{\theta_0+1-t\le x\le \theta_0+1+t}$$

  is an envelope for $\mathscr{F}$. Then, we have

$$\mathbb{P}F^2 \le \int_{\theta_0-1-t}^{\theta_0-1+t} p_{\theta_0}(x)\,\mathrm{d}x + \int_{\theta_0+1-t}^{\theta_0+1+t} p_{\theta_0}(x)\,\mathrm{d}x \le p_{\theta_0}(\theta_0)\cdot 4t \le C_{p_{\theta_0}} t < \infty$$

  for some constant $C_{p_{\theta_0}}$ depending on $p_{\theta_0}$. With the bracketing bound, for some constant $C > 0$,

$$\mathbb{E}\left[\sup_{f\in\mathscr{F}} \sqrt{n}|\mathbb{P}_n f - \mathbb{P}f|\right] \le C\cdot \|F\|_{L_2(\mathbb{P})} \int_0^1 \sqrt{\log N_{[\,]}(\mathscr{F}, L_2(\mathbb{P}), \epsilon\|F\|_{L_2(\mathbb{P})})}\,\mathrm{d}\epsilon.$$

  **Claim.** For all $\epsilon$, there exists some constant $C' > 0$ such that

$$N_{[\,]}(\mathscr{F}, L_2(\mathbb{P}), \epsilon) \le \left(\frac{1}{\epsilon}\right)^{C'} < \infty.$$

  With the above claim and $\|F\|_{L_2(\mathbb{P})} \le \sqrt{C_{p_{\theta_0}} t}$, the integral can be further bounded as

$$\int_0^1 \sqrt{\log N_{[\,]}(\mathscr{F}, L_2(\mathbb{P}), \epsilon\|F\|_{L_2(\mathbb{P})})}\,\mathrm{d}\epsilon \le \int_0^1 \sqrt{C' \log \frac{1}{\epsilon\|F\|_{L_2(\mathbb{P})}}}\,\mathrm{d}\epsilon < \infty,$$

  hence, there exists some constant $C > 0$ such that

$$\mathbb{E}\left[\sup_{f\in\mathscr{F}} |\mathbb{P}_n f - \mathbb{P}f|\right] \le C\sqrt{\frac{t}{n}}.$$

  This motivates us to define $\phi_n(t)$ as $C\sqrt{t/n}$.
- To check the sub-quadratic assumption, for all $n$ and $c > 1$, we have

$$\phi_n(ct) = C\sqrt{\frac{ct}{n}} = \sqrt{c}\cdot\left(C\sqrt{\frac{t}{n}}\right) = \sqrt{c}\cdot\phi_n$$

  hence the sub-quadratic assumption is satisfied with $\alpha = 1/2$.
- Consider the rate-determining equation $\phi_n(t) \le t^2$, for $t = \delta_n$,

$$\sqrt{t/n} \le t^2 \Leftrightarrow \sqrt{\delta_n/n} \le \delta_n^2 \Rightarrow 1/\sqrt{n} \le \delta_n^{3/2} \Rightarrow \delta_n \approx 1/n^{1/3}.$$

In all, we have $|\hat{\theta}_n - \theta_0| = O_p(n^{-1/3})$. ∎

Proposition 4.3.1 is not fully proven yet, since we only have the growth condition satisfied in a ball

around $\theta_0$, not for all $\theta \in \Theta$.

# Lecture 22: More Examples on Rate of Convergence

Before we extend Theorem 4.3.1 to handle the local growth condition, we note the following. 20 Oct. 9:00

**Remark.** The rate obtained from Theorem 4.3.1 is usually correct; on the other hand, the probability tail bound obtained from Theorem 4.3.1 is

$$\mathbb{P}(d(\hat{\theta}, \theta_0) > t\delta_n) \lesssim \frac{1}{t},$$

with $t = 2^M$ in the argument, which can be weak in the sense that it does not imply $\mathbb{E}[d(\hat{\theta}, \theta_0)] = O(\delta_n)$. Potentially, more sophisticated concentration arguments can be used.

**Remark.** The main step to apply Theorem 4.3.1 is to bound the expected supremum of localized empirical process, which can be hard.

Finally, as shown in some situations, we cannot expect the growth condition to hold for all $\theta \in \Theta$; instead, typically we only have $\theta \in B(\theta_0, u^*)$ for some $u^* \in \mathbb{R}$. In this case, a variation of Theorem 4.3.1 still holds [VW96, Theorem 3.2.5].

**Theorem 4.3.2.** For an $M$-estimation problem, assume the growth condition on $M$ holds for $\theta \in B(\theta_0, u^*)$ for some $u^*$, and the sub-quadratic assumption (with parameter $\alpha < 2$) and the rate-determining equation are valid for $\phi_n$'s arose from bounding the localized empirical process,

$$\mathbb{P}(d(\hat{\theta}_n, \theta_0) > 2^M \delta_n) \leq 4c \sum_{j > M} 2^{(\alpha-2)j}.$$

**Proof.** We again start by doing the peeling step, but this time consider

$$\mathbb{P}(d(\hat{\theta}, \theta_0) > 2^M \delta_n) \leq \sum_{\substack{j > M: \\ 2^j \delta_n \leq u^*}} \mathbb{P}(2^{j-1}\delta_n < d(\hat{\theta}, \theta_0) < 2^j \delta_n) + \mathbb{P}\left(d(\hat{\theta}, \theta_0) > \frac{u^*}{2}\right).$$

We then handle the first term as in Theorem 4.3.1, and show the second term goes to 0. ∎

## 4.3.3 More Examples

We see two more examples of using Theorem 4.3.2.

**Example** (Sample quantile). Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$ which has density $f$ w.r.t. Lebesgue measure. Moreover, for $0 < \tau < 1$, let

$$\rho_\tau(x) = \begin{cases} (\tau - 1), & \text{if } x < 0; \\ \tau x, & \text{if } x \geq 0, \end{cases}$$

and $m_\theta(x) = \rho_\theta(x - \theta)$ for all $\theta \in \mathbb{R}$, so

$$M(\theta) = \mathbb{E}\left[m_\theta(x)\right], \quad M_n(\theta) = \frac{1}{n}\sum_{i=1}^n \rho_\tau(x_i - \theta)$$

We see that $\theta_0 \coloneqq \arg\min_\theta M(\theta)$ is the $\tau^{\text{th}}$ quantile of $\mathbb{P}$, and let $\hat{\theta} \coloneqq \arg\min_\theta M_n(\theta)$ be the corresponding $M$-estimator. The rate of convergence is $|\hat{\theta}_n - \theta_0| = O_p(1/\sqrt{n})$.

**Proof.** To show the rate of convergence, consider the following.

- $|m_{\theta_1}(x) - m_{\theta_2}(x)| \le |\theta_1 - \theta_2|$, i.e., this is a Lipschitz parametric class.
- To show the growth condition, we need the following.

**Lemma 4.3.1.** For all $w, v \in \mathbb{R}$,

$$\rho_\tau(w-u) - \rho_\tau(w) = -v(\tau - \mathbb{1}_{w\le 0}) + \int_0^v [\mathbb{1}_{w\le z} - \mathbb{1}_{w\le 0}]\,\mathrm{d}z.$$

Then, we can show the growth condition satisfy in a neighborhood of $\theta_0$.

**Claim.** For $d(\theta,\theta_0) = |\theta-\theta_0|$, for $\theta$ in some neighborhood of $\theta_0$, $|\theta-\theta_0|^2 \lesssim M(\theta) - M(\theta_0)$.

**Proof.** By denoting $F$ as the CDF of $f$, we have

$$\begin{aligned} M(\theta_0+\delta) - M(\theta_0) &= \mathbb{E}\left[\rho_\tau(x-\theta_0-\delta) - \rho_\tau(x-\theta_0)\right] \\ &= \mathbb{E}\left[-\delta(\tau - \mathbb{1}_{x-\theta_0\le 0})\right] + \mathbb{E}\left[\int_0^\delta (\mathbb{1}_{x-\theta_0\le z} - \mathbb{1}_{x-\theta_0\le 0})\,\mathrm{d}z\right] \\ &= \int_0^\delta F(\theta_0+z) - F(\theta_0)\,\mathrm{d}z \end{aligned}$$

assume there exists a neighborhood of $\theta_0$ such that $f \ge L > 0$, then for some $\xi_z \in (0,\delta)$,

$$\ge \int_0^\delta f(\xi_z) z\,\mathrm{d}z \ge L\cdot\int_0^\delta z\,\mathrm{d}z = \frac{L\delta^2}{2},$$

i.e., in this neighborhood, $M$ grows quadratically in a neighborhood of $\theta_0$. ⊛

- To bound the localized empirical process, first note that $\hat\theta$ is consistent,[a] and consider $\mathscr{F} = \{m_\theta - m_{\theta_0} : |\theta-\theta_0| \le t\}$ where the localized empirical process is

$$\mathbb{E}\left[\sup_{\theta:\ |\theta-\theta_0|\le t} |(\mathbb{P}_n - \mathbb{P})m_\theta - (\mathbb{P}_n - \mathbb{P})m_{\theta_0}|\right].$$

To use the bracketing bound, we first see that $F(x) = t$ is an envelope with $\|F\|_{L_2(\mathbb{P})} = t$, hence the localized empirical process can be upper-bounded by

$$\frac{t}{\sqrt{n}}\int_0^1 \sqrt{\log N_{[\,]}(\mathscr{F}, L_2(\mathbb{P}), \epsilon t)}\,\mathrm{d}\epsilon \le \frac{t}{\sqrt{n}}\int_0^1 \sqrt{\log\left(1+\frac{4}{\epsilon}\right)}\,\mathrm{d}\epsilon,$$

i.e., the localized empirical process can be upper-bounded by $\phi_n(t) \approx t/\sqrt{n}$.

- It's evident that $\phi_n$'s satisfy the sub-quadratic assumption with $\alpha = 1$.
- By the rate-determining equation,

$$\delta_n/\sqrt{n} \approx \delta_n^2 \Rightarrow \delta_n = 1/\sqrt{n},$$

implying $|\hat\theta_n - \theta_0| = O_p(1/\sqrt{n})$.

⊛

[a] This should be proved beforehand, otherwise things doesn't make sense.

**Remark.** In this sample quantile example, the classical result for $\tau = 0.5$ shows that

$$\sqrt{n}(\hat\theta - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{4(f(\theta_0))^2}\right).$$

Another example is the high-dimensional linear regression.

**Example** (High-dimensional linear regression). Consider $Z = (Y, X_1, \dots, X_p) \in \mathbb{R}^{p+1}$ such that $Z_1, \dots, Z_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, and we want to predict $Y$ by $\beta^\top X$. Let $M(\beta) = \mathbb{E}\left[(Y - \beta^\top X)^2\right]$ with

$$\beta^* = \underset{\beta\colon \|\beta\|_1 \le L}{\arg\min}\ M(\beta)$$

for some $L$, and let $M_n(\beta) = \frac{1}{n}\sum_{i=1}^n (Y^i - \beta^\top X^i)^2$ with

$$\hat{\beta} = \underset{\beta\colon \|\beta\|_1 \le L}{\arg\min}\ M_n(\beta).$$

**Intuition.** We want a *sparse* $\beta^*$, which is achieved by controlling $L$.

**Note.** We're not assuming the underlying $\mathbb{P}$ to be linear.

The main question of interest for the high dimensional linear regression is the following.

**Problem** (Persistency). How large can $L = L(n,p)$ be so $M(\hat{\beta}) - M(\beta^*) \to 0$ as $n, p \to \infty$?

**Answer.** With some assumptions, Theorem 4.3.3 shows that $L \lesssim \sqrt[4]{\log p/n}$. ⊛

**Theorem 4.3.3.** Under the setup of high-dimensional linear regression, let $Y = X_0$ and define $F(Z) = \max_{0\le j,k\le p} |X_j X_k - \mathbb{E}[X_j X_k]|$. Assume further that $\mathbb{E}\left[F^2(Z)\right] < \infty$, then

$$M(\hat{\beta}) - M(\beta^*) = O_p\left(L, \sqrt[4]{\frac{\log p}{n}}\right).$$

**Proof.** From the basic inequality, $M(\hat{\beta}) - M(\beta^*) \le 2\sup_{\beta\colon \|\beta\|_1 \le L} |M_n(\beta) - M(\beta)|$. By letting

$$\gamma = (-1, \beta)^\top, \quad \Sigma = \left(\mathbb{E}\left[X_j^1 X_k^1\right]\right)_{j,k=0,\dots,p}, \quad \Sigma_n = \left(\frac{1}{n}\sum_{i=1}^n X_j^i X_n^i\right)_{j,k=0,\dots,p},$$

we may then write $M(\beta) = \gamma^\top \Sigma \gamma$ and $M_n(\beta) = \gamma^\top \Sigma_n \gamma$. Hence,

$$\sup_{\beta\colon \|\beta\|_1 \le L} |M_n(\beta) - M(\beta)| = |\gamma^\top \Sigma_n \gamma - \gamma^\top \Sigma \gamma| \le \|\gamma\|_1^2 \cdot \|\Sigma_n - \Sigma\|_\infty \le (1+L)^2 \|\Sigma_n - \Sigma\|_\infty,$$

which implies

$$\sup_{\beta\colon \|\beta\|_1 \le L} \mathbb{P}(M(\hat{\beta}) - M(\beta^*) > \epsilon) \le P((1+L)^2\|\Sigma_n - \Sigma\|_\infty > \epsilon) \le \frac{(1+L)^2 \mathbb{E}\left[\|\Sigma_n - \Sigma\|_\infty\right]}{\epsilon}$$

by Markov's inequality. Finally, we observe that

$$\mathbb{E}\left[\|\Sigma_n - \Sigma\|_\infty\right] = \mathbb{E}\left[\sup_{f \in \mathscr{F}} |\mathbb{P}_n f - \mathbb{P} f|\right]$$

where $\mathscr{F} = \{f_{jk}\colon 0 \le j,k \le p\}$ with $f_{jk} = X_j X_k - \mathbb{E}[X_j X_k]$. Now, $F$ is clearly an envelope with $\|F\|_{L_2(\mathbb{P})} < \infty$, and by defining $\epsilon$-brackets to be $[f_{j,k} \pm \epsilon/2]$ for all $j, k = 0, \dots, p$, we have

$$N_{[\,]}(\mathscr{F}, L_2(\mathbb{P}), \epsilon) \le (p+1)^2.$$

By the bracketing bound, $\mathbb{E}\left[\|\Sigma_n - \Sigma\|_\infty\right] \leq \frac{1}{\sqrt{n}}\|F\|_{L_2(\mathbb{P})}\sqrt{2\log(p+1)}$, i.e.,

$$M(\hat{\beta}) - M(\beta^*) \leq \frac{(1+L)^2}{\sqrt{n}}\|F\|_{L_2(\mathbb{P})}\sqrt{2\log(p+1)}.$$

In order to let this goes to 0, we require $L \lesssim \sqrt[4]{n/\log p}$, which finally implies

$$M(\hat{\beta}) - M(\beta^*) = O_p\left(L, \sqrt[4]{\frac{\log p}{n}}\right).$$

■

**Remark.** Observe that in the above proof, we do not need to localize the empirical process, i.e., the bracketing bound can be used for any empirical process induced by finite class.

# Chapter 5

# Fixed Design Non-Parametric Regression

## Lecture 23: Smooth Constrained Least Square

In this chapter, we're going to study the *fixed design non-parametric regression*, a more advanced application based on $M$-estimations. Here, fixed design refers to the case that when an empirical risk minimization problem has data $X_i$'s given are "fixed", i.e., the only randomness comes from $Y_i$'s.[1] 23 Oct. 9:00

## 5.1 Smooth Constrained Least Square

In this section, we will see an example of constrained least square called smooth constrained least square. This helps us prepare for developing a more unified theory towards the former problem.

### 5.1.1 Prediction and Estimation: Statistical Learning v.s. $M$-Estimation

Since we're going to consider regression problems, i.e., empirical risk minimizations, with the tools developed for $M$-estimations, it's natural to compare "prediction" and "estimation". Recall the following.

> **As previously seen** (Prediction). Given $(X_1, Y_1), \dots, (X_n, Y_n) \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, let $f^*(x) = \mathbb{E}[Y \mid X = x]$, i.e.,
>
> $$f^* = \underset{f \text{ measurable}}{\arg\min}\ \mathbb{E}_{(X,Y)}\left[(Y - f(X))^2\right].$$
>
> More generally, given a class of function $\mathscr{F}$, we can define the predictor
>
> $$f_{\mathscr{F}} = \underset{f \in \mathscr{F}}{\arg\min}\, \mathbb{E}\left[(Y - f(X))^2\right].$$
>
> For any prediction function $f : \chi \to \mathbb{R}$, its excess risk is
>
> $$\mathbb{E}\left[(Y - f(X))^2\right] - \inf_{h \in \mathscr{F}} \mathbb{E}\left[(Y - h(X))^2\right].$$

We now ask, how does this "prediction" approach relate to the problem of "estimation", e.g., $M$-estimation? Observe that we can write

$$\begin{aligned}
&\mathbb{E}\left[(Y - f(X))^2\right] - \inf_{h \in \mathscr{F}} \mathbb{E}\left[(Y - h(X))^2\right] \\
&= \mathbb{E}\left[(Y - f(X) \pm f^*(X))^2\right] - \inf_{h \in \mathscr{F}} \mathbb{E}\left[(Y - h(X) \pm f^*(X))^2\right] \\
&= \mathbb{E}\left[(f(X) - f^*(X))^2\right] - \inf_{h \in \mathscr{F}} \mathbb{E}\left[(h(X) - f^*(X))^2\right] && \text{cross terms are 0} \\
&= \|f - f^*\|_{L_2(\mathbb{P})}^2 - \inf_{h \in \mathscr{F}} \|h - f^*\|_{L_2(\mathbb{P})}^2,
\end{aligned}$$

[1] In contrast, there is also *random design*, where $X_i$'s are also sampled from a distribution.

where the last equation is the estimation error in the traditional statistics terminology. Looking at the last line, we see that "prediction" and "estimation" are the same in this content.

**Intuition.** Predicting $\mathbb{E}[Y \mid X = x]$ given $x$ is equivalent to estimating $f^*$ in the view of minimizing excess risk and $L_2(\mathbb{P})$ difference, respectively.

### 5.1.2 Oracle Inequality

If we change $f$ in the above bound to a predictor $\hat{f}_n$ (depending on the training data), we have the so-called *oracle inequality*

$$\mathbb{E}\left[(Y - \hat{f}_n(X))^2\right] - \inf_{h \in \mathscr{F}} \mathbb{E}\left[(Y - h(X))^2\right] = \|\hat{f}_n - f^*\|_{L_2(\mathbb{P})}^2 - \inf_{h \in \mathscr{F}} \|h - f^*\|_{L_2(\mathbb{P})}^2 \tag{5.1}$$

There are two standard scenarios regarding bounding the oracle inequality.

**Definition 5.1.1** (Well specified). Given a function class $\mathscr{F}$, an empirical risk minimization is *well specified* if $f^* \in \mathscr{F}$, i.e., $\mathbb{P}$ is such that $f(x) = \mathbb{E}[Y \mid X = x] \in \mathscr{F}$.

In the well specified case, the excess risk of $\hat{f}_n$ is just $\|\hat{f}_n - f^*\|_{L_2(\mathbb{P})}^2$ since $\inf_{h \in \mathscr{F}} \|h - f^*\|_{L_2(\mathbb{P})}^2 = 0$, i.e., the estimation error in $L_2(\mathbb{P})$. This is what we usually assume in the traditional statistics. On the other hand, we can also consider the contrary.

**Definition 5.1.2** (Misspecified). Given a function class $\mathscr{F}$, an empirical risk minimization is *misspecified* if $f^* \notin \mathscr{F}$, i.e., $\mathbb{P}$ is such that $f(x) = \mathbb{E}[Y \mid X = x] \notin \mathscr{F}$.

In the misspecified case, we need to bound the oracle inequality entirely.

### 5.1.3 Smooth Constrained Least Square

First, consider the following fixed-design problem with data generated at the fixed grid $x \in \{1/n, \dots, 1\}$.

**Problem 5.1.1** (Smooth constrained least square). Let $\mathscr{F} = \{f \colon [0,1] \to [-1,1] \text{ 1-Lipschitz}\}$, and consider $\epsilon_1, \dots, \epsilon_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ such that for a fixed $f^* \in \mathscr{F}$, for all $i = 1, \dots, n$,

$$y_i = f^*(i/n) + \epsilon_i$$

Then, the problem of *smooth constrained least square* is to find the least square estimate

$$\hat{f}_n = \arg\min_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(i/n))^2 .$$

**Notation** (Sequence model). The *sequence model* refers to the noise $\epsilon_i$ being $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$.

We may use our techniques in $M$-estimations due to the following.

**Remark.** $\hat{f}_n$ is an $M$-estimator.

**Proof.** We first observe that from the definition of $y_i$,[a]

$$\hat{f}_n = \arg\max_{f \in \mathscr{F}} \frac{2}{n} \sum_{i=1}^n \epsilon_i (f(i/n) - f^*(i/n)) - \frac{1}{n} \sum_{i=1}^n (f(i/n) - f^*(i/n))^2 ,$$

which motivates us to define $M_n(f), M(f) \colon \mathscr{F} \to \mathbb{R}$ such that

$$M_n(f) = \frac{2}{n}\sum_{i=1}^n \epsilon_i \left(f(i/n) - f^*(i/n)\right) - \frac{1}{n}\sum_{i=1}^n \left(f(i/n) - f^*(i/n)\right)^2$$

and

$$M(f) = \mathbb{E}\left[M_n(f)\right] = -\frac{1}{n}\sum_{i=1}^n \left(f(i/n) - f^*(i/n)\right)^2.$$

We see that $f^* = \arg\max_{f\in\mathscr{F}} M(f)$. ⊛

[a] Note that the term $-\frac{1}{n}\sum_{i=1}^n \epsilon_i^2$ is omitted since it doesn't depend on $f \in \mathscr{F}$.

Since $\hat{f}_n$ is an $M$-estimator, we want to get the rate. We need to verify the following.

**Claim.** The growth condition is satisfied for some $d$.

**Proof.** By choosing the canonical $d$, i.e.,

$$d(f, f^*) \coloneqq \sqrt{M(f^*) - M(f)} = \sqrt{\frac{1}{n}\sum_{i=1}^n \left(f(i/n) - f^*(i/n)\right)^2} = L_2(\mathbb{P}_n)(f, f^*),$$

where $\mathbb{P}_n$ is the empirical measure on the grid $\{1/n, 2/n, \dots, 1\}$, we see that it satisfies the growth condition automatically. ⊛

Next, we bound the localized empirical process for the smooth constrained least square, i.e.,

$$\mathbb{E}\left[\sup_{f\in\mathscr{F}\colon d(f,f^*)\le\delta} (M_n - M)(f) - (M_n - M)(f^*)\right].$$

**Claim.** This localized empirical process is bounded by $\phi_n(\delta) = c\sqrt{\delta/n}$ for some $c > 0$.

**Proof.** We first observe that

$$\mathbb{E}\left[\sup_{\substack{f\in\mathscr{F}:\\ d(f,f^*)\le\delta}} (M_n - M)(f) - (M_n - M)(f^*)\right] = 2\cdot\mathbb{E}\left[\sup_{\substack{f\in\mathscr{F}:\\ d(f,f^*)\le\delta}} \frac{1}{n}\sum_{i=1}^n \epsilon_i \left(f(i/n) - f^*(i/n)\right)\right].$$

For $f \in \mathscr{F}$, define $X_f = \frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i \left(f(i/n) - f^*(i/n)\right)$, so for $f, g \in \mathscr{F}$,

$$X_f - X_g = \frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i \left(f(i/n) - g(i/n)\right) \sim \mathcal{N}(0, \sigma^2 d^2(f, g)),$$

which implies that $X_f$ is a sub-Gaussian process with

$$\mathbb{P}(|X_f - X_g| \ge u) \le \exp\left(-\frac{u^2}{2d^2(f,g)}\right)$$

when $\sigma^2 = 1$.[a] So, by Dudley's entropy integral bound, for some $c' > 0$,

$$\mathbb{E}\left[\sup_{\substack{f\in\mathscr{F}:\\ d(f,f^*)\le\delta}} \frac{1}{n}\sum_{i=1}^n \epsilon_i \left(f(i/n) - f^*(i/n)\right)\right] \le \frac{12}{\sqrt{n}}\int_0^\delta \sqrt{\log N(\mathscr{F}, d, \epsilon)}\,\mathrm{d}\epsilon \le \frac{12}{\sqrt{n}}\int_0^\delta \sqrt{\frac{c_1}{\epsilon}}\,\mathrm{d}\epsilon = c'\sqrt{\frac{\delta}{n}}$$

since $N(\mathscr{F}, d, \epsilon) \le N(\mathscr{F}, \|\cdot\|_\infty, \epsilon) \le \exp(c_1/\epsilon)$ from Theorem 3.3.1. Setting $c = 2c'$ suffices. ⊛

[a] Without this assumption, we just have $\sigma^2$ in the final bound, which is still a sub-Gaussian process.

Then, we verify the sub-quadratic assumption for such $\phi_n$'s.

**Claim.** $\phi_n$'s satisfy the sub-quadratic assumption for $\alpha = 1/2$.

**Proof.** Since $\phi_n(c\delta) = \sqrt{c} \cdot c\sqrt{\delta/n} = \sqrt{c} \cdot \phi_n(\delta)$. ⊛

With all these, we can finally solve the rate-determining equation, which is

$$\phi_n(\delta) \approx \delta^2 \Rightarrow \sqrt{\frac{\delta}{n}} \approx \delta^2 \Rightarrow \delta \approx n^{-1/3}.$$

In all, we have the rate of convergence

$$d(\hat{f}_n, f^*) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_n(i/n) - f^*(i/n)\right)^2} = O_p(n^{-1/3}),$$

or more specifically, from Theorem 4.3.2,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_n(i/n) - f^*(i/n)\right)^2 > 2^{2M} n^{-2/3}\right) \leq c \cdot 2^{-M}.$$

**Remark.** The rate of convergence in mean-square error (i.e., w.r.t. $d^2 = L_2^2(\mathbb{P})$) for 1-Lipschitz regression is $n^{-2/3}$. Moreover, the "min-max rate" worst case over all $f^* \in \mathscr{F}$ is $n^{-2/3}$, hence the smooth constrained least square estimator is "min-max rate optimal". These terms will make sense in the next remark.

More generally, if we assume that $f^* \in \mathcal{S}_\alpha$, we can also solve the smooth constrained least square over $\mathcal{S}_\alpha$ as follows.

**Theorem 5.1.1.** Consider the smooth constrained least square problem over $\mathcal{S}_\alpha$, the rate of convergence of $d(\hat{f}_n, f^*)$ for the canonical $d$ is $O_p(n^{-\frac{\alpha}{2\alpha+1}})$ for $\alpha > 1/2$.

**Proof.** From the same calculation, the corresponding localized empirical process is bounded by

$$\begin{aligned}\mathbb{E}\left[\sup_{\substack{f\in\mathcal{S}_\alpha:\\ d(f,f^*)\leq\delta}} \frac{1}{n}\sum_{i=1}^{n}\epsilon_i\left(f(i/n) - f^*(i/n)\right)\right] &\leq \frac{c}{\sqrt{n}}\int_0^\delta \sqrt{\log N(\mathcal{S}_\alpha, d, \epsilon)}\,\mathrm{d}\epsilon \\ &\leq \frac{c}{\sqrt{n}}\int_0^\delta \sqrt{\log N(\mathcal{S}_\alpha, \|\cdot\|_\infty, \epsilon)}\,\mathrm{d}\epsilon \leq \frac{c}{\sqrt{n}}\int_0^\delta \left(\frac{1}{\epsilon}\right)^{\frac{1}{2\alpha}}\mathrm{d}\epsilon\end{aligned}$$

from Theorem 3.3.1. Since $\alpha > 1/2$, this bound gives $\lesssim \frac{1}{\sqrt{n}}\delta^{1-1/2\alpha}$, so we can take

$$\phi_n(\delta) := \frac{\delta^{1-\frac{1}{2\alpha}}}{\sqrt{n}}.$$

We see that the sub-quadratic assumption is satisfied since $\phi_n(c\delta) = c^{1-\frac{1}{2\alpha}}\phi_n(\delta)$ with $1 - 1/2\alpha < 2$. Solving the rate-determining equation, we have

$$\phi_n(\delta_n) \approx \delta_n^2 \Rightarrow \frac{\delta_n^{-\frac{1}{2\alpha}}}{\sqrt{n}} \approx \delta_n \Rightarrow \delta_n \approx n^{-\frac{\alpha}{2\alpha+1}}.$$

■

It's usually more natural to consider the rate for $d^2(\hat{f}_n, f^*) = L_2^2(\mathbb{P}_n)(\hat{f}_n, f^*) = O_p(n^{-\frac{2\alpha}{2\alpha+1}})$.

**Note.** When $\alpha < 1/2$, we need to use the modified version of Dudley's entropy integral bound. In this case, we will get a slower rate than $n^{-2\alpha/2\alpha+1}$ w.r.t. $d^2$.

**Remark** (Min-max rate optimal). $n^{-2\alpha/2\alpha+1}$ is *min-max rate optimal*, i.e., it's the "right" rate.

**Proof.** We just showed that $\mathbb{E}[L_2^2(\mathbb{P})(\hat{f}_n, f^*)] \le cn^{-2\alpha/(2\alpha+1)}$, and it's known that the *min-max rate* is lower-bounded by

$$\inf_{\hat{f}_n} \sup_{f \in \mathscr{F}} \mathbb{E}\left[L_2^2(\mathbb{P})(\hat{f}_n, f^*)\right] \ge cn^{-\frac{2\alpha}{2\alpha+1}}.$$

Hence, the rate for the smooth constrained least square estimator is *min-max rate optimal* over $\mathcal{S}_\alpha$ for $\alpha > 1/2$. However, it's not known whether the same conclusion holds for $\alpha < 1/2$. ⊛

We do not know $\alpha$. A natural question is the following.

**Problem.** Can we adaptively get $O(n^{-2\alpha/2\alpha+1})$ rate without knowing $\alpha$?

**Answer.** Yes. There are several ways to do this [Tsy08]. ⊛

# Lecture 24: Contraction Principle for Rademacher Complexity

## 5.1.4 Contraction Principle

25 Oct. 9:00

Before delving into the general constrained least square, we take a detour to see another way of bounding the excess risk.

**As previously seen.** Given $\mathscr{F} = \{f : \chi \to \mathscr{Y}\}$ and a loss $\ell(f(x), y)$, the excess risk of $\hat{f} \in \mathscr{F}$ is

$$\mathbb{E}[L(\hat{f})] - \inf_{f \in \mathscr{F}} L(f) = \mathbb{E}[\ell(\hat{f}(X), Y)] - \inf_{f \in \mathscr{F}} \mathbb{E}\left[\ell(f(X), Y)\right]$$

where $\hat{f}$ is the empirical risk minimizer $\hat{f} = \arg\min_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^n \ell(f(x_i), y_i)$.

We know that from Lemma 3.1.1 and symmetrization, the excess risk is bounded by

$$\mathbb{E}\left[\sup_{f \in \mathscr{F}} \left(\mathbb{E}\left[\ell(f(X), Y)\right] - \frac{1}{n}\sum_{i=1}^n \ell(f(x_i), y_i)\right)\right] \le 2 \cdot \mathbb{E}\left[\sup_{f \in \mathscr{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i \ell(f(x_i), y_i)\right],$$

i.e., two times of the Rademacher complexity (actually the Rademacher width) of the set

$$\ell \circ \mathscr{F}\big((x_1, y_1), \dots, (x_n, y_n)\big) \coloneqq \{\ell(f(x_1), y_1), \dots, \ell(f(x_n), y_n)\}_{f \in \mathscr{F}}.$$

We would like to further upper-bound this Rademacher complexity by the Rademacher complexity of $\mathscr{F}$ only. This can be done if $\ell$ is Lipschitz in the first argument for every fixed second argument.

**Theorem 5.1.2** (Contraction principle). Suppose we have $\Theta \subseteq \mathbb{R}^n$ with the Rademacher width $R(\Theta) = \mathbb{E}\left[\sup_{\theta \in \Theta} \langle \epsilon, \theta \rangle\right]$. Let $\phi_i : \mathbb{R} \to \mathbb{R}$ be 1-Lipschitz for $i = 1, \dots, n$, and let

$$\phi \circ \Theta \coloneqq \{(\phi_1(\theta_1), \phi_2(\theta_2), \dots, \phi_n(\theta_n)) : \theta \in \Theta\}.$$

Then, $R(\phi \circ \Theta) \le R(\Theta)$.

**Proof.** Let $\phi \circ \theta = (\phi_1(\theta_1), \phi_2(\theta_2), \dots, \phi_n(\theta_n))$, then

$$R(\phi \circ \Theta) = \mathbb{E}_{\epsilon_1, \dots, \epsilon_{n-1}}\left[\mathbb{E}_{\epsilon_n}\left[\sup_{\theta \in \Theta} \langle \phi \circ \theta, \epsilon \rangle\right]\right].$$

Condition on $\epsilon_1, \dots, \epsilon_{n-1}$, we have

$$\begin{aligned}
&\mathbb{E}_{\epsilon_n}\left[\sup_{\theta\in\Theta}\langle\phi\circ\theta,\epsilon\rangle\right] \\
&= \frac{1}{2}\left[\sup_{\theta\in\Theta}\left(\langle(\phi\circ\theta)_{1:\,n-1},\epsilon_{1:\,n-1}\rangle + \phi_n(\theta_n)\right) + \sup_{\theta'\in\Theta}\left(\langle(\phi\circ\theta')_{1:\,n-1},\epsilon_{1:\,n-1}\rangle - \phi_n(\theta'_n)\right)\right] \\
&= \frac{1}{2}\left[\sup_{\theta,\theta'\in\Theta}\left(\langle(\phi\circ\theta)_{1:\,n-1},\epsilon_{1:\,n-1}\rangle + \langle(\phi\circ\theta')_{1:\,n-1},\epsilon_{1:\,n-1}\rangle + \phi_n(\theta_n) - \phi_n(\theta'_n)\right)\right]
\end{aligned}$$

sine $\phi_n(\theta_n) - \phi_n(\theta'_n) \le |\theta_n - \theta'_n|$,

$$\begin{aligned}
&\le \frac{1}{2}\left[\sup_{\theta,\theta'\in\Theta}\left(\langle(\phi\circ\theta)_{1:\,n-1},\epsilon_{1:\,n-1}\rangle + \langle(\phi\circ\theta')_{1:\,n-1},\epsilon_{1:\,n-1}\rangle + |\theta_n - \theta'_n|\right)\right] \\
&= \frac{1}{2}\left[\sup_{\theta,\theta'\in\Theta}\left(\langle(\phi\circ\theta)_{1:\,n-1},\epsilon_{1:\,n-1}\rangle + \langle(\phi\circ\theta')_{1:\,n-1},\epsilon_{1:\,n-1}\rangle + \theta_n - \theta'_n\right)\right] \qquad \text{symmetry of } \theta,\theta' \\
&= \frac{1}{2}\left[\sup_{\theta\in\Theta}\left(\langle(\phi\circ\theta)_{1:\,n-1},\epsilon_{1:\,n-1}\rangle + \theta_n\right) + \sup_{\theta'\in\Theta}\left(\langle(\phi\circ\theta')_{1:\,n-1},\epsilon_{1:\,n-1}\rangle - \theta'_n\right)\right] \\
&= \frac{1}{2}\left[\sup_{\theta\in\Theta}\left(\langle(\phi\circ\theta)_{1:\,n-1},\epsilon_{1:\,n-1}\rangle + \theta_n\right) + \sup_{\theta\in\Theta}\left(\langle(\phi\circ\theta)_{1:\,n-1},\epsilon_{1:\,n-1}\rangle - \theta'_n\right)\right] \\
&= \mathbb{E}_{\epsilon_n}\left[\sup_{\theta\in\Theta}\langle(\phi\circ\theta)_{1:\,n-1},\epsilon_{1:\,n-1}\rangle + \theta_n\epsilon_n\right].
\end{aligned}$$

We see that we have got rid of $\phi_n$! Now, by considering iterating this method for all $i$, and taking the expectation over all $\epsilon_i$'s, we can then get rid of $\phi$ entirely, and hence get $R(\Theta)$ finally. ■

**Corollary 5.1.1.** The same conclusion of contraction principle holds for $\phi_i$'s being $L$-Lipschitz, specifically, $R(\phi\circ\Theta) \le L\cdot R(\Theta)$.

Now, going back to bounding excess risk, we let $\phi_i(x) := \ell(x, y_i)$.[2] If these $\phi_i$'s are $L$-Lipschitz, the contraction principle with $\Theta = \{(f(x_1),\dots,f(x_n))\}_{f\in\mathscr{F}} \subseteq \mathbb{R}^n$ gives

$$\mathbb{E}\left[\sup_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^n \epsilon_i \ell(f(x_i), y_i)\right] \le L\cdot\mathbb{E}\left[\sup_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)\right]$$

where $\ell(f(x_i), y_i) = \phi_i(f(x_i))$.

**Example.** Square loss is Lipschitz if $\mathscr{F}$ is uniformly bounded and $\mathscr{Y}$ is also uniformly bounded.

## 5.2 Well Specified Constrained Least Square

Now, we consider the generalization of smooth constrained least square in the well specified case.

**Problem 5.2.1** (Constrained least square)**.** Given a function class $\mathscr{F}$ on $\chi$, let $x_1,\dots,x_n \in \chi$ and $y_i = f^*(x_i) + \epsilon_i$ where $\epsilon_1,\dots,\epsilon_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma^2)$[a] and assume that $f^* \in \mathscr{F}$. Then, the *constrained least square* aims to estimate $f^*$ via

$$\hat{f} = \arg\min_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2.$$

[a] This is known as *Gaussian sequence model*.

[2] The second argument is fixed at $y_i$ for each $\phi_i$ as we noted before.

**Note.** The constrained least square generalizes smooth constrained least square by a general given $\mathscr{F}$ with data given not on the fixed grid, but just some fixed points in $\chi$.

Next, we will see that "local"[3] Gaussian complexity of $\mathscr{F}$ around $f^*$ determines the rate of convergence of $\hat{f}$. Specifically, the goal is to give non-asymptotic bounds on

$$\|\hat{f} - f^*\|_{L_2(\mathbb{P}_n)}^2 = \frac{1}{n}\sum_{i=1}^n \left(\hat{f}(x_i) - f^*(x_i)\right)^2$$

for $\mathbb{P}_n = \{x_1, \dots, x_n\}$.

**Note.** In contrast, in random design, we assume $x_1, \dots, x_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, and the goal is to give bound on

$$\|\hat{f} - f^*\|_{L_2(\mathbb{P})}^2 = \mathbb{E}_X\left[(\hat{f}(X) - f^*(X))^2\right].$$

**Example** (Image denoising). In the case of fixed design, a typical example is image denoising where $x_i = (i/n, j/n)$ are the pixels.

# Lecture 25: Constrained Least Square

## 5.2.1 Basic Inequality

27 Oct. 9:00

First, observe that we can think of $\mathscr{F} \subseteq \mathbb{R}^n$, i.e., $f^* \in \mathbb{R}^n$, $y \in \mathbb{R}^n$, and $\epsilon \in \mathbb{R}^n$ since all we care about is the values on the fixed points $x_1, \dots, x_n$. In this case, $y = f^* + \epsilon$, and the goal is to find an estimation vector $\hat{f} \in \mathbb{R}^n$, and we want to find a bound on

$$\frac{1}{n}\|\hat{f} - f^*\|^2.$$

Let's see some interesting examples of $\mathscr{F}$.

**Notation** (Risk). The *risk* is defined as

$$R(\hat{f}, f^*) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (\hat{f}_i - f_i^*)^2\right] = \mathbb{E}\left[\frac{1}{n}\|\hat{f} - f^*\|^2\right].$$

**Example** (Linear regression). When $f^* = X\beta^*$ for some $X_{n\times d}$, we can regard $\mathscr{F} = \mathcal{C}(X_{n\times d})$, where $\mathcal{C}$ denotes the column space. In this case, $\hat{f} = X\hat{\beta}$ where $\hat{\beta}$ is the ordinary least square estimator with the risk being

$$\mathbb{E}\left[\frac{1}{n}\|X\hat{\beta} - X\beta^*\|^2\right] = \frac{\operatorname{rank}(X)}{n}\cdot\sigma^2.$$

**Example** (Smoothness constraint). If $\theta_i^* = f^*(i/n)$ for $f^* \in S_\alpha$, i.e., the smooth constrained least square, the risk has order $O(n^{-\frac{2\alpha}{2\alpha+1}})$ for $\alpha > 1/2$ as shown in Theorem 5.1.1.

**Example** (Constrained Lasso). $\mathscr{F} = \{X\beta \colon \|\beta\|_1 \le L\}$

**Example** (Isotonic regression). $\mathscr{F} = \{f \colon f_1 \le f_2 \le \cdots \le f_n\}$

[3] Recall the localization we did before, i.e., the localized empirical process. We will do similar things here.

**Example** (Low rank). $\mathscr{F} = \{f \in \mathbb{R}^{n\times n} \colon \operatorname{rank}(f) = k\}$ for some small $k$.

**Example.** $\mathscr{F} = \{f \colon \sum_{i=1}^{n} |f_{i+1} - f_i| \le V\}$.

**Example.** $\mathscr{F} = \{f \colon \sum_{i=1}^{m}\sum_{j=1}^{m} |f_{i+1,j} - f_{i,j}| + |f_{i,j+1}, f_{i,j}| \le V\}$.

To summarize, the constraints come from $\mathscr{F}$ in the constrained least square. Now, to start bounding $\|\hat{f} - f^*\|$, we first see a basic inequality: since $f^* \in \mathscr{F}$ and $\hat{f}$ is the minimizer of $\|y - \hat{f}\|^2/n$,

$$\|y - \hat{f}\|^2 \le \|y - f^*\|^2.$$

With $y = f^* + \epsilon$, the above is equivalent to

$$\begin{aligned}
&\|f^* + \epsilon - \hat{f}\|^2 \le \|\epsilon\|^2 \\
\Leftrightarrow\ &\langle f^* + \epsilon - \hat{f}, f^* + \epsilon - \hat{f}\rangle \le \langle \epsilon, \epsilon\rangle \\
\Leftrightarrow\ &\langle f^* + \epsilon - \hat{f}, f^* + \epsilon - \hat{f}\rangle - \langle f^* + \epsilon - \hat{f}, \epsilon\rangle - \langle \epsilon, f^* - \hat{f}\rangle \le \langle \epsilon, \epsilon\rangle - \langle f^* + \epsilon - \hat{f}, \epsilon\rangle - \langle \epsilon, f^* - \hat{f}\rangle \\
\Leftrightarrow\ &\langle f^* - \hat{f}, f^* - \hat{f}\rangle \le \langle \hat{f} - f^*, \epsilon\rangle - \langle \epsilon, f^* - \hat{f}\rangle \\
\Leftrightarrow\ &\langle \hat{f} - f^*, \hat{f} - f^*\rangle \le 2\langle \epsilon, \hat{f} - f^*\rangle \\
\Leftrightarrow\ &\|\hat{f} - f^*\|^2 \le 2\langle \epsilon, \hat{f} - f^*\rangle.
\end{aligned}$$

We call this the *basic inequality*

$$\|\hat{f} - f^*\|^2 \le 2\langle \epsilon, \hat{f} - f^*\rangle \tag{5.2}$$

Consider a toy example.

**Example.** Let $\mathscr{F} = \mathcal{C}(X_{n\times d})$, then $\hat{f} = P_X y$ where $P_X$ is the project matrix. From the basic inequality, $\|X\hat{\beta} - X\beta^*\|^2 \le 2\langle \epsilon, X\hat{\beta} - X\beta^*\rangle$. Dividing both sides by $\|X\hat{\beta} - X\beta^*\|$ leads to

$$\|X\hat{\beta} - X\beta^*\| \le 2\left\langle \epsilon, \frac{X\hat{\beta} - X\beta^*}{\|X\hat{\beta} - X\beta^*\|}\right\rangle \le 2 \cdot \sup_{v = X\beta - X\beta^*} \left\langle \epsilon, \frac{v}{\|v\|}\right\rangle = 2 \sup_{\substack{v\in\mathcal{C}(X)\\ \|v\|=1}} \langle \epsilon, v\rangle = 2\|P_X\epsilon\|,$$

so we have $\|X\hat{\beta} - X\beta^*\| \le 2\|P_X\epsilon\|$.[a]

[a] In this specific example, the factor 2 is superfluous since $X\hat{\beta} - X\beta^* = P_X y = X\beta^* + P_X\epsilon - X\beta^*$. Furthermore, we have $\mathbb{E}\left[\|P_X\epsilon\|^2\right] \sim \chi^2_{\operatorname{rank}(X)}$.

### 5.2.2 Localized Gaussian Width

Now, going back to the general case, we note the following.

**Note** (Compact $\mathscr{F}$). If $\mathscr{F}$ is compact, then the basic inequality gives us an upper-bound as

$$\|\hat{f} - f^*\|^2 \le 2 \sup_{f\in\mathscr{F}} \langle \epsilon, f - f^*\rangle.$$

**Intuition.** This upper-bound depends on the complexity of the neighborhood around $f^*$ within $\mathscr{F}$.

By mimicking the above calculation, we can replace the basic inequality by

$$\|\hat{f} - f^*\| \le 2 \sup_{f\in\mathscr{F}} \left\langle \epsilon, \frac{f - f^*}{\|f - f^*\|}\right\rangle.$$

To simplify the notation, let $\mathscr{F}^* = \{f - f^* \colon f \in \mathscr{F}\}$, i.e., we center $\mathscr{F}$ by $f^*$, then the bounds becomes

$$\|\hat{f} - f^*\| \le 2 \sup_{f\in\mathscr{F}^*} \left\langle \epsilon, \frac{f}{\|f\|}\right\rangle.$$

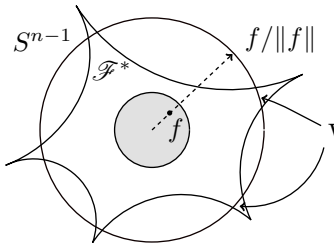
We see that this supremum depends on the set $\{f/\|f\|\colon f \in \mathscr{F}^*\}$, and actually it's just the (scaled) Gaussian width of $\{f/\|f\|\colon f \in \mathscr{F}^*\}$.

**Note.** The set $\{f/\|f\|\}_{f\in\mathscr{F}^*}$ is the (sub)set of unit vectors, i.e., $\{f/\|f\|\}_{f\in\mathscr{F}^*} \subseteq S^{n-1}$. If this set is the full sphere $S^{n-1}$, then the supremum is just $2\|\epsilon\| = O_p(\sqrt{n})$. In this case,

$$\frac{1}{n}\|\hat{f} - f^*\|^2 = O(1),$$

i.e., the risk doesn't go to 0, which is not good, indicating that $\mathscr{F}$ is too large.

Hence, we need $\mathscr{F}^*$ to be significantly "smaller" than $S^{n-1}$ in the Gaussian width sense; like linear subspace with rank $d$. However, we observe that as long as there's a smaller ball inside $\mathscr{F}^*$, this rescaling will force us to consider all $S^{n-1}$, leading to non-convergence risk.



What we really want to consider

**Intuition.** We should take advantage of the fact that at a larger scale, we don't have a smaller sphere.

This suggests us to look instead at that for some $t > 0$, $\sup_{f\in\mathscr{F}^*\colon \|f\|\geq t} \langle \epsilon, f/\|f\|\rangle$. To do this, consider

$$\begin{aligned}
\|\hat{f} - f^*\| &= \mathbb{1}_{\|\hat{f}-f^*\|\leq t} \cdot \|\hat{f} - f^*\| + \mathbb{1}_{\|\hat{f}-f^*\|>t} \cdot \|\hat{f} - f^*\| \\
&\leq t + \mathbb{1}_{\|\hat{f}-f^*\|>t} \cdot \|\hat{f} - f^*\| \\
&\leq t + 2\left\langle \epsilon, \frac{\hat{f} - f^*}{\|\hat{f} - f^*\|} \right\rangle \cdot \mathbb{1}_{\|\hat{f}-f^*\|>t} \\
&\leq t + 2 \sup_{f\in\mathscr{F}^*\colon \|f\|\geq t} \left\langle \epsilon, \frac{f}{\|f\|} \right\rangle.
\end{aligned}$$

**Intuition.** This is like the peeling step we did before, but now we do this for a single-scale $t$.

To control the intersections between $\mathscr{F}^*$ and $S^{n-1}$, a good assumption is the following.

**Definition 5.2.1** (Star-shaped). A class of functions $\mathscr{H}$ is *star-shaped* around 0 if for all $h \in \mathscr{H}$, $\lambda h \in \mathscr{H}$ for $\lambda \in [0, 1]$.

**Example** (Convex $\mathscr{H}$). If $\mathscr{H}$ is convex and contains 0, then it's star-shaped.

Now, if $\mathscr{F}^*$ is star-shaped around 0, for any $f \in \mathscr{F}^*$ with $\|f\| > t$, there exists $u \in \mathscr{F}^*$ such that $u = t \cdot f/\|f\|$ with $\|u\| = t$. Hence,

$$\left\langle \epsilon, \frac{f}{\|f\|} \right\rangle = \frac{1}{t}\langle \epsilon, u\rangle,$$

i.e., under the assumption that $\mathscr{F}^*$ is star-shaped, we can rewrite the bound as

$$\|\hat{f} - f^*\| \leq t + \frac{2}{t} \sup_{\substack{u\in\mathscr{F}^* \\ \|u\|\leq t}} \langle \epsilon, u\rangle \leq t + \frac{2}{t} \sup_{\substack{f\in\mathscr{F}^* \\ \|f\|\leq t}} \langle \epsilon, f\rangle. \tag{5.3}$$

This is the same as the localization we did before, where $\sup_{f\in\mathscr{F}^*\colon \|f\|\leq t}\langle \epsilon, f\rangle$ is just the supremum of the localized empirical process.

**Intuition.** To bound this further, we will use concentration for the supremum.

First, denote

$$Z(t) = \sup_{\substack{f \in \mathscr{F}^* \\ \|f\| \le t}} \langle \epsilon, f \rangle, \quad G(t) = \mathbb{E}\left[Z(t)\right],$$

so Equation 5.3 becomes

$$\|\hat{f} - f^*\| \le t + \frac{2}{t} Z(t). \tag{5.4}$$

Digress from bounding $Z(t)$ for now, we note that the definition of $G(t)$ reminds us the (scaled) Gaussian width for the set $\mathscr{F}^*$, where the difference lies in the norm constraint in $Z(t)$, i.e., $\|f\| \le t$. This suggests the following definition.

**Definition 5.2.2** (Localized Gaussian width). Let $g_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Then the *localized Gaussian width* of a set $A \subseteq \mathbb{R}^n$ of some $t > 0$ is defined as

$$\mathrm{GW}_n(A,t) = \mathbb{E}\left[\sup_{a \in A:\ \|a\| \le t} \frac{1}{n} \langle g, a \rangle\right].$$

We note that we can just use Gaussian width to represent localized Gaussian width:

**Remark.** Since $\mathrm{GW}_n(A,t) = \mathrm{GW}_n(A \cap B(0,t))$, so localized Gaussian width is just a shorthand notation we don't really need to introduce.

**Note.** If $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma^2)$ with $\sigma^2 = 1$, then $G(t) = \mathbb{E}\left[\sup_{f \in \mathscr{F}^*:\ \|f\| \le t} \langle \epsilon, f \rangle\right] = n\,\mathrm{GW}_n(\mathscr{F}^*, t)$.

Back to bounding $Z(t)$. By showing $Z(t) \approx G(t)$ with high probability, the bound becomes $t + 2G(t)/t$. Balancing these two terms and looking for the minimum, we get the tightest upper-bound. Since balancing $t$ and $2G(t)/t$ means setting these two terms to $G(t) = t^2/2$, this suggests the following.

**Definition 5.2.3** (Critical radius). The *critical radius* for a constrained least square is defined as

$$t^* = \inf_{t > 0} \{t : G(t) \le t^2/2\}.$$

In this case, the critical radius will minimize the bound, hence $\|\hat{f} - f^*\| = O(t^*)$.

# Lecture 26: Rate of Constrained Least Square

Before we proceed, we note the following. 30 Oct. 9:00

**Remark.** $t^*$ is well-defined, i.e., there exists $t$ such that $G(t) \le t^2/2$.

**Proof.** From Cauchy-Schwarz,

$$G(t) \le \mathbb{E}\left[\|\epsilon\| \cdot t\right] \le \left(\mathbb{E}\left[\|\epsilon\|^2\right]\right)^{1/2} t \le \sqrt{n} \cdot t.$$

Hence, $G$ is bounded by a linear function of $t$, so there exists $t$ such that $G(t) \le t^2/2$. ⊛

**Remark.** The rate of $t^* \to \infty$ with $n \to \infty$ is typically that $t^*/n \to 0$.

We first show some property of $G(t)/t$.

**Lemma 5.2.1.** If $\mathscr{F}^*$ is star-shaped, then the map $t \mapsto G(t)/t$ is non-increasing.

**Proof.** Take $t^* \le t' \le t$ and any $f \in \mathscr{F}^*$ such that $\|f\| \le t$. By the star-shaped assumption on $\mathscr{F}^*$, $t' \cdot f/t \in \mathscr{F}^*$. Note that $\|t'/t \cdot f\| \le t'$, which implies

$$Z(t) = \sup_{\substack{f \in \mathscr{F}^* \\ \|f\| \le t}} \langle \epsilon, f \rangle = \frac{t}{t'} \cdot \sup_{\substack{f \in \mathscr{F}^* \\ \|f\| \le t}} \left\langle \epsilon, \frac{t'}{t} \cdot f \right\rangle \le \frac{t}{t'} \cdot \sup_{\substack{f \in \mathscr{F}^* \\ \|f\| \le t'}} \langle \epsilon, f \rangle = \frac{t}{t'} Z(t').$$

Hence, $Z(t) \le t/t' \cdot Z(t')$, i.e., $Z(t)/t \le Z(t')/t'$. Taking the expectation gives the result. ■

Now, observe the following.

**Corollary 5.2.1.** For any $t \ge t^*$, $G(t) \le t^2/2$.

**Proof.** Let $t \ge t^*$, we have

$$G(t) = t \cdot \frac{G(t)}{t} \le t \cdot \frac{G(t^*)}{t^*} \le t \cdot \frac{t^*}{2} \le \frac{t^2}{2}$$

where the first and second inequality follows from Lemma 5.2.1 and Definition 5.2.3. ■

Now, we return to bound Equation 5.4 with the concentration argument. We have the following.

**Theorem 5.2.1** (Gaussian concentration for Lipschitz functions). Let $\epsilon_1, \dots, \epsilon_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ and $\phi \colon \mathbb{R}^n \to \mathbb{R}$ be $L$-Lipschitz w.r.t. the Euclidean norm, i.e., $|\phi(x) - \phi(y)| \le L\|x - y\|_2$. Then for every $t \ge 0$,

$$\mathbb{P}\left(\phi(\epsilon) - \mathbb{E}\left[\phi(\epsilon)\right] > t\right) \le e^{-\frac{t^2}{2L^2\sigma^2}}.$$

Theorem 5.2.1 means that $\phi(\epsilon)$ has the same tail decay as $\mathcal{N}(0, L)$. Now, to apply Theorem 5.2.1, we need to show that $Z(t)$ is indeed Lipschitz.

**Lemma 5.2.2.** $Z(t)$ is a $t$-Lipschitz function of $\epsilon$.

**Proof.** Let $Z(t) = Z(t, \epsilon)$, then

$$\phi(\epsilon) - \phi(\epsilon') = Z(t, \epsilon) - Z(t, \epsilon') = \sup_{\substack{f \in \mathscr{F}^* \\ \|f\| \le t}} \langle \epsilon, f \rangle - \sup_{\substack{f \in \mathscr{F}^* \\ \|f\| \le t}} \langle \epsilon', f \rangle.$$

Let $\hat{f} = \arg\sup_{f \in \mathscr{F}^* \|f\| \le t} \langle \epsilon, f \rangle$, we further have

$$\phi(\epsilon) - \phi(\epsilon') \le \langle \epsilon, \hat{f} \rangle - \langle \epsilon', \hat{f} \rangle = \langle \epsilon - \epsilon', \hat{f} \rangle \le \|\epsilon - \epsilon'\| \|\hat{f}\| \le t\|\epsilon - \epsilon'\|$$

from Cauchy-Schwarz. The other side is the same. ■

Hence, by combining Theorem 5.2.1 and Lemma 5.2.2, we have the following.

**Corollary 5.2.2.** For every $u, t \ge 0$,

$$\mathbb{P}\left(Z(t) - G(t) \ge u\right) \le \exp\left(-\frac{u^2}{2t^2\sigma^2}\right).$$

**Note.** Set $u = t^2$ in Corollary 5.2.2, we have $\mathbb{P}\left(Z(t) \ge G(t) + t^2\right) \le e^{-t^2/2\sigma^2}$.

Hence, for all $t \ge t^*$, from Corollary 5.2.2, with probability at least $1 - e^{-t^2/2\sigma^2}$,

$$Z(t) \le G(t) + t^2 \le \frac{t^2}{2} + t^2 = \frac{3}{2}t^2$$

with Corollary 5.2.1. In all, with probability at least $1 - e^{-t^2/2\sigma^2}$, the bound Equation 5.4 becomes

$$\|\hat{f} - f^*\| \le t + \frac{2}{t} Z(t) \le 4t,$$

i.e., we have shown that for all $t \geq t^*$,

$$\mathbb{P}\left(\|\hat{f} - f^*\| \leq 4t\right) \geq 1 - e^{-\frac{t^2}{2\sigma^2}}.$$

Putting this into a theorem, we have the following.

**Theorem 5.2.2** (Non-asymptotic bound on constrained least square)**.** Consider the constrained least square over $\mathscr{F}$ such that $\mathscr{F}^*$ is star-shaped, and let $t^*$ be the critical radius. Then for all $s \geq 1$,

$$\mathbb{P}\left(\|\hat{f} - f^*\| \leq 4st^*\right) \geq 1 - e^{-\frac{(st^*)^2}{2\sigma^2}}.$$

**Proof.** We let $t \geq t^*$ to be $t \coloneqq st^*$ for some $s \geq 1$, iterative the above argument gives the result. ■

The expectation version of Theorem 5.2.2 is the following.

**Remark.** For some constant $C > 0$, $\mathbb{E}[\|\hat{f} - f^*\|^2] \leq C(\sigma^2 t^{*2} + 1)$.

We make several important notes.

**Note.** Theorem 5.2.2 says that $t^*$ determines the mean square error between $\hat{f}$ and $f^*$, where the explicit non-asymptotic one-sided tail bound states that the tail decay is Gaussian.

**Note.** Theorem 5.2.2 can be generalized to bounded $\epsilon_1, \dots, \epsilon_n$.

**Proof.** Then the main step of proving Theorem 5.2.2 is Theorem 5.2.1, which does not hold in this case. However, a similar concentration result holds for the convex and Lipschitz functions of $\epsilon$. Hence, showing $Z(t)$ is convex as a function of $\epsilon$ and with Lemma 5.2.2, i.e., $Z(t)$ is $t$-Lipschitz suffices. ⊛

**Remark** (Convex $\mathscr{F}$ [Cha14])**.** If $\mathscr{F}$ is convex,[a] a two-sided tail bound holds under a different definition of critical radius, specifically, $t^* = \arg\max_t G(t) - t^2/2$. In particular,

$$t^{*2} - C \cdot t^{*3/2} \leq \mathbb{E}\left[\|\hat{f} - f^*\|^2\right] \leq t^{*2} + C \cdot t^{*3/2}.$$

[a] Which is stronger than being star-shaped as noted before.

# Lecture 27: Another Approach of Localization

## 5.2.3 Offset Gaussian Rademacher Complexity

1 Nov. 9:00

We now see another approach of bounding the basic inequality $\|\hat{f} - f^*\|^2 \leq 2\langle \epsilon, \hat{f} - f^* \rangle$ given a constrained least square problem over $\mathscr{F}$. First, instead of dividing both sides by $\|\hat{f} - f^*\|$, we first double both sides and rearrange, we have

$$\|\hat{f} - f^*\|^2 \leq 4\langle \epsilon, \hat{f} - f^* \rangle - \|\hat{f} - f^*\|^2 \leq \sup_{f \in \mathscr{F}} 4\langle \epsilon, f - f^* \rangle - \|f - f^*\|^2.$$

Hence, our goal now is to bound this so-called offset Gaussian Rademacher complexity of $\mathscr{F}$ around $f^*$.

**Definition 5.2.4** (Offset Gaussian Rademacher complexity)**.** Consider the constrained least square problem over $\mathscr{F}$ with $f^* \in \mathscr{F}$. Then the *offset Gaussian Rademacher complexity* of $\mathscr{F}$ around $f^*$ is defined as

$$\sup_{f \in \mathscr{F}} 4\langle \epsilon, f - f^* \rangle - \|f - f^*\|^2.$$

**Example.** Consider $\mathscr{F} = \{X\beta \colon \beta \in \mathbb{R}^d\}$, where $X$ is a fixed $n \times d$ design matrix. Then one can show that the corresponding offset Gaussian Rademacher complexity is

$$\sup_{\beta \in \mathbb{R}^d} \langle \epsilon, X\beta \rangle - \|X\beta\|^2 \approx \operatorname{rank}(X^\top X).$$

Note that $\epsilon$ is random, so a natural goal now is to consider the expected offset Gaussian Rademacher complexity and show that it concentrates around its expectation. Indeed, we can show this as follows.

**Theorem 5.2.3.** Consider the constrained least square over $\mathscr{F}$ such that $\mathscr{F}^*$ is star-shaped, and let $t^*$ be the critical radius. Then for all $s > 0$, the offset Gaussian Rademacher complexity satisfies

$$\mathbb{P}\left( \sup_{f \in \mathscr{F}^*} 4\langle \epsilon, f \rangle - \|f\|^2 \le 4(t^* + s)^2 \right) \ge 1 - e^{-\frac{s^2}{2\sigma^2}}.$$

Consequently, $\mathbb{P}(\|\hat{f} - f^*\|^2 \le 4(t^* + s)^2) \ge 1 - e^{-\frac{s^2}{2\sigma^2}}$.

**Proof.** From Corollary 5.2.2, let $u = st^*$, we have

$$\mathbb{P}\left(Z(t^*) \le G(t^*) + st^*\right) \ge 1 - e^{-\frac{s^2}{2\sigma^2}}.$$

Denote the event of $Z(t^*) \le G(t^*) + st^*$ by $E$, take any $f \in \mathscr{F}^*$.

- If $\|f\| \le t^*$: then $4\langle \epsilon, f \rangle - \|f\|^2 \le 4\langle \epsilon, f \rangle \le 4Z(t^*)$. On the event of $E$, we further have

  $$4\langle \epsilon, f \rangle - \|f\|^2 \le 4(G(t^*) + st^*) \le 4\left(\frac{t^{*2}}{2} + st^*\right) \le 4\left(\frac{t^{*2}}{2} + st^* + \frac{s^2}{2}\right) = 2(t^* + s)^2,$$

  where the second inequality follows from Corollary 5.2.1.

- If $\|f\| > t^*$: let $\gamma = t^*/\|f\| < 1$, then $\|\gamma f\| = t^*$, hence we have

  $$4\langle \epsilon, f \rangle - \|f\|^2 = \frac{4}{\gamma}\langle \epsilon, \gamma f \rangle - \frac{t^{*2}}{\gamma^2} \le \frac{4}{\gamma} Z(t^*) - \frac{t^{*2}}{\gamma^2} = \frac{4t^*}{\gamma}\frac{Z(t^*)}{t^*} - \frac{t^{*2}}{\gamma^2} \le \left(2\frac{Z(t^*)}{t^*}\right)^2$$

  from $2ab - b^2 \le a^2$ with $a = 2Z(t^*)/t^*$ and $b = t^*/\gamma$. On the event $E$, we further have

  $$4\langle \epsilon, f \rangle - \|f\|^2 \le 4\left(\frac{Z(t^*)}{t^*}\right)^2 \le 4\frac{(G(t^*) + st^*)^2}{t^{*2}} \le 4\left(\frac{t^{*2}/2 + st^*}{t^*}\right)^2 \le 4(t^* + s)^2$$

  where the third inequality again follows from Corollary 5.2.1.

By combining two bounds, we have the desired result. ∎

We see that by using either localized Gaussian width or offset Gaussian Rademacher complexity, we successfully get the rate of convergence for the risk of constrained least square.

**Remark.** Comparing Theorem 5.2.3 and Theorem 5.2.2, we see that

- Theorem 5.2.3: $\|\hat{f} - f^*\|^2$ after $t^{*2}$ has a tail with exponentially decay.
- Theorem 5.2.2: $\|\hat{f} - f^*\|$ after $t^*$ has a tail with Gaussian decay.

We note that Theorem 5.2.3 directly generalizes to constants other than 4.

**Corollary 5.2.3.** Consider the constrained least square over $\mathscr{F}$ such that $\mathscr{F}^*$ is star-shaped, and let

$t^*$ be the critical radius. Then for all $s, c > 0$, the offset Gaussian Rademacher complexity satisfies

$$\mathbb{P}\left(\sup_{f\in\mathscr{F}^*} c\langle\epsilon, f\rangle - \|f\|^2 \le \frac{c^2}{2}(t^*+s)^2\right) \ge 1 - e^{-\frac{s^2}{2\sigma^2}}.$$

### 5.2.4 Rate of Convergence for Constrained Least Square

We note that in order to apply either Theorem 5.2.3 or Theorem 5.2.2, the critical step is to first find $t^*$, which involves bounding $G(t)$ and equating it with $t^2/2$. First, recall the following.

**As previously seen.** Bounding $G(t)$ is equivalent to bounding the localized Gaussian width since

$$G(t) = \mathbb{E}\left[\sup_{f\in\mathscr{F}^*:\,\|f\|\le t}\langle\epsilon, f\rangle\right] = n\cdot \mathrm{GW}(\mathscr{F}-f^*, t).$$

Essentially, it is an expectation of the supremum of the localized empirical process, so we may apply the bracketing bound or Dudley's entropy bound, which in terms depends on the covering number.

Let's now see some examples of $\mathscr{F}$ of the constrained least square problem and their corresponding localized Gaussian width. Consider the fixed grid setup as in the smooth constrained least square: given a function class $\mathscr{F}$ from $\mathbb{R}$ to $\mathbb{R}$, define the set $\mathscr{F}$[4] to be

$$\mathscr{F} = \left\{\big(f(1/n), f(2/n), \dots, f(1)\big) \in \mathbb{R}^n : f\in\mathscr{F}\right\}.$$

We note the following.

**Remark.** One can often replace $\mathrm{GW}(\mathscr{F}-f^*, t)$ by $\mathrm{GW}(\mathscr{F}-\mathscr{F}, t)$ without getting a worse rate.

**Proof.** We see that these two localized Gaussian widths correspond to the following supremum

$$\mathrm{GW}(\mathscr{F}-f^*,t) \Leftrightarrow \sup_{f\in\mathscr{F}} f - f^*, \quad \text{and} \quad \mathrm{GW}(\mathscr{F}-\mathscr{F},t) \Leftrightarrow \sup_{f,g\in\mathscr{F}} f-g$$

The latter gives the worst case rate since the supremum is taken "uniformly" over all $g\in\mathscr{F}$. However, the local geometry of $\mathscr{F}$ is usually quite "uniform", so the rate doesn't blow up.

Another reason is that since our machinery requires $\mathscr{F}^*$ to be star-shaped, but if it's not, as long as $\mathscr{F}$ is, then $\mathscr{F}-\mathscr{F}$ is also star-shaped, which is another reason to look at $\mathscr{F}-\mathscr{F}$. ⊛

We can then look at some examples of $\mathscr{F}$ with their corresponding log-covering number.

**Example** (Hölder smooth functions). Consider $\mathscr{F} = \mathcal{S}_\alpha$,[a] with $\alpha > 1/2$, for some $c > 0$,

$$\log N(\mathscr{F}-\mathscr{F}, \|\cdot\|_2/\sqrt{n}, \epsilon) \le c\left(\frac{1}{\epsilon}\right)^{1/\alpha}.$$

The rate of $\mathbb{E}[\|\hat{f}-f^*\|^2]/\sqrt{n}$ is $n^{-\frac{\alpha}{2\alpha+1}}$.

[a] Note that actually $\mathscr{F}$ is a subset in $\mathbb{R}^n$ evaluated on the grid.

**Proof.** If $\mathscr{F}$ is $L$-Lipschitz, then $\mathscr{F}-\mathscr{F}$ is $2L$-Lipschitz. Similar result holds for $\mathcal{S}_\alpha$ for $\alpha > 1/2$, hence Theorem 3.3.1 applies. Then from Theorem 5.1.1 with the $1/\sqrt{n}$-scaling, we get the rate. ⊛

It's worth noting that another function class different from $\mathcal{S}_\alpha$ that are studied are shape-constrained function classes. We now see another example where we consider $\mathscr{F}-\mathscr{F}$ instead.

**Example** (Unimodal functions). Consider the class of *unimodal functions* defined as

$$\mathscr{F} = \{(f_1,\dots,f_n)\in\mathbb{R}^n : \exists i:\, f_1\le f_2\le\cdots\le f_i\ge f_{i+1}\ge\cdots\ge f_n\},$$

[4] We're abusing the notation here: we iteratively define the "set" $\mathscr{F}$ by the function class $\mathscr{F}$.

which is non-convex but star-shaped around 0. Then for some $c > 0$,

$$\log N(\mathscr{F}, \|\cdot\|_2/\sqrt{n}, \epsilon) \le \frac{c}{\epsilon}$$

with $\mathbb{E}[\|\hat{f} - f^*\|^2]/\sqrt{n} \le n^{-2/3}$.

**Proof.** Observe that $\mathscr{F}$ is a finite union of convex sets. If we take any $f^* \in \mathscr{F}$, $\mathscr{F} - f^*$ is not necessarily star-shaped, so we consider $\mathscr{F} - \mathscr{F}$ instead since at least it itself is star-shaped. ⊛

Lastly, let's consider the class of isotonic functions, where the sequence $\{f_i\}$ is increasing.

**Example** (Bounded isotonic functions)**.** Consider the class of *bounded isotonic functions* defined as $\mathscr{F} = \{(f_1, \dots, f_n) \in \mathbb{R}^n : 0 \le f_1 \le f_2 \le \cdots \le f_n \le 1\}$. Then for some $c > 0$, we have

$$\log N(\mathscr{F}, \|\cdot\|_2/\sqrt{n}, \epsilon) \le \frac{c}{\epsilon}.$$

This is similar as 1-Lipschitz functions, so the rate of $\mathbb{E}[\|\hat{f} - f^*\|^2]/\sqrt{n}$ is just $1/n^{2/3}$.

**Example** (Unbounded isotonic functions)**.** If we instead consider the class of *unbounded isotonic functions* defined as $\mathscr{F} = \{(f_1, \dots, f_n) : f_1 \le f_2 \le \cdots \le f_n\}$, i.e., $\hat{f}$ will have no tuning parameters. Then it's known that for $v^* = (f_n^* - f_1^*)$,

$$\mathrm{GW}(\mathscr{F} - f^*, t) \le C n^{1/4} v^{*1/2} t^{1/2}$$

with $\mathbb{E}[\|\hat{f} - f^*\|^2] \le (v^*/n)^{2/3}$.

**Proof.** First, we note that obtaining such a bound is possible even if $\mathscr{F}$ is non-compact,[a] i.e., $\mathrm{GW}(\mathscr{F}^*, t)$ is finite even if $\mathscr{F}$ is unbounded. Moreover, if we accept the bound on GW, then equating this with $t^2$ gives

$$t^* \approx v_n^{*1/3} n^{1/6} \Rightarrow \mathbb{E}\left[\|\hat{f} - f^*\|^2\right] \lesssim t^{*2}/n \approx (v^*/n)^{2/3}$$

where we need to divide by $n$ due to $G(t) = n\,\mathrm{GW}(\mathscr{F}^*, t)$. ⊛

[a] Recall the previous remark: the set we want to evaluate the Gaussian width is the intersection of $\mathscr{F}$ with some bounded ball, which is compact.

In general, if $\mathscr{F}$ is an isotonic function class with $f^*$ being a smooth increasing function, then the mean square error will be of the order $n^{-2/3}$; and if $f^*$ is a piecewise constant function with $k$ pieces, then the rate of the least squared estimators will be of order $\frac{k}{n}\log n$, which is faster.

**Intuition.** The above means that the geometry of $\mathscr{F}$ is different around different $f^*$.

To conclude, the above intuition is true in general since $G(t)$ (and hence the localized Gaussian width) is a geometric quantity around $f^*$ as a function of $t$, hence, this depends on the choice of $f^*$.

# Lecture 28: Misspecified Non-Parametric Regression

## 5.3 Misspecified Constrained Least Square

8 Nov. 9:00

So far, we assume that $f^* \in \mathscr{F}$, i.e., we're assuming the well-specified case. Now, we consider $f^* \notin \mathscr{F}$ with the goal of finding a non-asymptotic bound on the oracle inequality

$$\|\hat{f} - f^*\|_{L_2(\mathbb{P})}^2 - \inf_{f \in \mathscr{F}} \|f - f^*\|_{L_2(\mathbb{P})}^2$$

for an estimator $\hat{f}$ (potentially depends on $n$).

### 5.3.1 Convex Function Class

We start by considering the case when $\mathscr{F}$ is convex. In the well specified case, recall the following.

> **As previously seen.** If $f^* \in \mathscr{F}$ and $\hat{f}$ is the minimizer of $\frac{1}{n}\|y - \hat{f}\|^2$, $\|y - \hat{f}\|^2 \le \|y - f^*\|^2$. This is the first step of deriving the basic inequality.

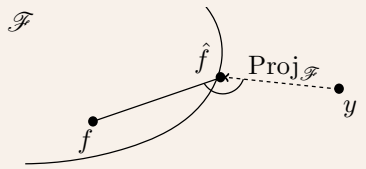
However, if $f^* \notin \mathscr{F}$, the basic inequality cannot be used since the first step already breaks down. In this case, we may instead consider $\text{Proj}_{\mathscr{F}}(f^*)$ to force $f^*$ "lies" in $\mathscr{F}$. Note the following fact.

> **Note.** If $\mathscr{F}$ is closed and convex, then the project is uniquely defined.

> **Proof.** For any $y \in \mathbb{R}^n$, $\hat{f} = \arg\min_{f \in \mathscr{F}} \|y - f\|^2$ uniquely exists, and the least square solution corresponds to the projection of $y$ to $\mathscr{F}$.[a] ⊛
>
> [a] This holds for infinite dimensional vector space $\mathscr{F}$ too from functional analysis. See note.

Using this, we might derive the following mimicking the basic inequality.

**Lemma 5.3.1** (Obtuse angle lemma). Let $\mathscr{F}$ be a closed and convex set, then for any $y \in \mathbb{R}^n$, $\hat{f} = \text{Proj}_{\mathscr{F}}(y)$ if and only if for every $f \in \mathscr{F}$, $\langle y - \hat{f}, f - \hat{f}\rangle \le 0$.




**Proof.** For every $f \in \mathscr{F}$, consider the squared distance between $y$ and the line between $\hat{f}$ and $f$

$$d\colon [0,1] \to \mathbb{R}, \quad d(\lambda) = \|y - (\lambda\hat{f} + (1-\lambda)f)\|^2.$$

Since $\mathscr{F}$ is closed and convex, $d(\lambda)$ is minimized as $\lambda \to 1$ if and only if $\hat{f} = \text{Proj}_{\mathscr{F}}(y)$, i.e.,

$$\frac{\mathrm{d}d(\lambda)}{\mathrm{d}\lambda}\bigg|_{\lambda \to 1} = 2(y - (\lambda\hat{f} + (1-\lambda)f))^\top (f - \hat{f})\Big|_{\lambda \to 1} = 2(y - \hat{f})^\top (f - \hat{f}),$$

with the fact that we're dealing with a constrained optimization, the above is not $= 0$, but $\le 0$. ■

> **Remark.** The obtuse angle lemma is the characterizing property of $\hat{f}$, which corresponds to the KKT condition of the corresponding optimization problem.

We can now bound $\|y - \hat{f}\|^2$ for $\hat{f} = \text{Proj}_{\mathscr{F}}(y)$ by considering for any $f \in \mathscr{F}$,

$$\|y - f\|^2 = \|y - \hat{f}\|^2 + \|\hat{f} - f\|^2 + 2\langle y - \hat{f}, \hat{f} - f\rangle \ge \|y - \hat{f}\|^2 + \|\hat{f} - f\|^2,$$

since $2\langle y - \hat{f}, \hat{f} - f\rangle \ge 0$ from the obtuse angle lemma, we conclude

$$\|y - \hat{f}\|^2 \le \|y - f\|^2 - \|\hat{f} - f\|^2.$$

This leads to the improved version of the basic inequality by plugging in $y = f^* + \epsilon$,

$$\|f^* - \hat{f}\|^2 + \|\epsilon\|^2 + 2\langle \epsilon, f^* - \hat{f}\rangle \le \|f^* - f\|^2 + \|\epsilon\|^2 + 2\langle \epsilon, f^* - f\rangle - \|\hat{f} - f\|^2.$$

Rearranging, we get

$$\|\hat{f} - f^*\|^2 - \|f - f^*\|^2 \le 2\langle \epsilon, \hat{f} - f\rangle - \|\hat{f} - f\|^2.$$

Since this is true for all $f \in \mathscr{F}$, so we can use this for $\overline{f} = \text{Proj}_{\mathscr{F}}(f^*)$, hence

$$\|\hat{f} - f^*\|^2 - \|\overline{f} - f^*\|^2 \le 2\langle \epsilon, \hat{f} - \overline{f}\rangle - \|\hat{f} - \overline{f}\|^2$$

with $\|\overline{f} - f^*\|^2 = \inf_{f\in\mathscr{F}}\|f - f^*\|^2$, we get the *misspecified basic inequality* for convex $\mathscr{F}$

$$\|\hat{f} - f^*\|^2 \le \inf_{f\in\mathscr{F}}\|f - f^*\|^2 + 2\langle\epsilon, \hat{f} - \overline{f}\rangle - \|\hat{f} - \overline{f}\|^2. \tag{5.5}$$

To provide a uniform upper-bound, we may take a supremum over the last two terms, which yields

$$\|\hat{f} - f^*\|^2 \le \inf_{f\in\mathscr{F}}\|f - f^*\|^2 + \sup_{f\in\mathscr{F}-\overline{f}} 2\langle\epsilon, f\rangle - \|f\|^2.$$

We see that the supremum is just the (skewed) offset Gaussian Rademacher complexity of $\mathscr{F} - \overline{f}$, and from Corollary 5.2.3, it's $\lesssim t^{*2}$ with high probability using the developed critical radius machinery.

> **Remark** (Well specified). If $f^* \in \mathscr{F}$, i.e., back to the well specified case, then $\inf_{f\in\mathscr{F}}\|f - f^*\|^2 = 0$, and we just get back to the previous result.[a]
>
> [a] Actually, now this is even with a better constant 2 than 4.

### 5.3.2 Non-Convex Function Class

If $\mathscr{F}$ is not convex, we do not have the KKT condition, nor the misspecified basic inequality just derived. In this case, least square, i.e., the projection approach can be "suboptimal" in terms of bounding $\|\hat{f} - f^*\|^2$. To understand what do we mean by "suboptimal" here, consider the following.

> **Intuition.** If we look at the oracle inequality, then we're essentially considering
>
> $$\|\hat{f} - f^*\|^2 \le c \cdot \inf_{f\in\mathscr{F}}\|f - f^*\|^2 + \text{bound of oracle inequality},$$
>
> where we're interested in $c = 1$.[a]
>
> [a] We call this the *sharp* oracle inequality.

In this point of view, it's known that taking the least square over the convex hull of $\mathscr{F}$ can be suboptimal when $c = 1$. To overcome this, consider a "two-step estimation". Let

$$\hat{g} = \arg\min_{f\in\mathscr{F}}\|y - f\|^2 \quad \text{and} \quad \hat{f} = \arg\min_{f\in\text{Star}(\hat{g},\mathscr{F})}\|y - f\|^2,$$

where $\text{Star}(\hat{g}, \mathscr{F}) = \{\lambda\hat{g} + (1-\lambda)f \colon f \in \mathscr{F}, 0 \le \lambda \le 1\}$. More generally, consider the following.

> **Definition 5.3.1** (Star set). Given a set $A$ and an element $x$,[a] the *star set* of $A$ w.r.t. $x$ is defined as
>
> $$\text{Star}(x, A) \coloneqq \{\lambda x + (1-\lambda)a \colon a \in A, 0 \le \lambda \le 1\}.$$
>
> More generally, given two sets $A, B$, we define $\text{Star}(A, B) \coloneqq \{\lambda a + (1-\lambda)b \colon a \in A, b \in B, 0 \le \lambda \le 1\}$.
>
> [a] Not necessary in $A$.

This two-step estimator $\hat{f}$ is also called *star-estimator* due to the use of star set.

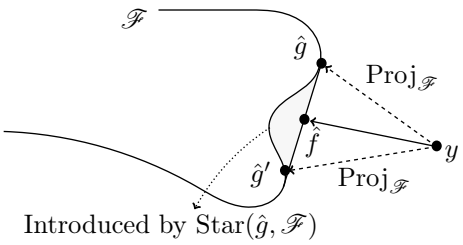



Figure 5.1: Illustration of two-step estimator $\hat{f}$. Note that $\hat{g}$ might not be unique.

The upshot is that one can show the following.

**Lemma 5.3.2** (Improved basic inequality). In the misspecified constrained least square, for all $f \in \mathscr{F}$,

$$\|y - \hat{f}\|^2 \le \|y - f\|^2 - \frac{1}{18}\|\hat{f} - f\|^2$$

**Remark.** The improved basic inequality is like the Pythagorean inequality for non-convex sets.

As a consequence, plug $y = f^* + \epsilon$ in the improved basic inequality, for all $f \in \mathscr{F}$ we have

$$\|\hat{f} - f^*\|^2 - \|f - f^*\|^2 \le 2\langle \epsilon, \hat{f} - f\rangle - \frac{1}{18}\|\hat{f} - f\|^2.$$

Let $f = \overline{f} \coloneqq \mathrm{Proj}_{\mathscr{F}}(f^*)$, and again note that $\|\overline{f} - f^*\|^2 = \inf_{f\in\mathscr{F}}\|f - f^*\|^2$, we finally get

$$\begin{aligned}\|\hat{f} - f^*\|^2 &\le \|\overline{f} - f^*\|^2 + 2\langle \epsilon, \hat{f} - \overline{f}\rangle - \frac{1}{18}\|\hat{f} - \overline{f}\|^2 \\ &\le \inf_{f\in\mathscr{F}}\|f - f^*\|^2 + \sup_{f\in \mathrm{Star}(\mathscr{F},\mathscr{F})-\overline{f}} 2\langle \epsilon, f\rangle - \frac{1}{18}\|f\|^2.\end{aligned}$$

We see that the supremum is again the (non-evenly skewed) offset Gaussian Rademacher complexity of the convex hull of $\mathscr{F}$[5] minus $\overline{f}$.

**Remark.** The complexities of $\mathrm{Star}(\mathscr{F},\mathscr{F}) - \overline{f}$ and $\mathscr{F}$ itself are often of the same order.

# Lecture 29: Convex Penalized Estimators

## 5.4 Penalized Least Square

10 Nov. 9:00

Now, instead of constraining $\mathscr{F}$ in the constrained least square, we may start by considering $\mathscr{F} = \mathbb{R}^n$ (i.e., no constraint) but use a different objective to "penalize" some estimators, or equivalently, to favor a particular class of estimators. Specifically, consider the following problem.

**Problem 5.4.1** (Penalized least square). Let $\theta^* \in \mathbb{R}^n$ and $y = \theta^* + \epsilon$ where $\epsilon_1, \dots, \epsilon_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$. Then, given a penalty function $f\colon \mathbb{R}^n \to \mathbb{R}$ and a tuning parameter $\lambda \in \mathbb{R}$, the *penalized least square* aims to estimate $\theta^*$ via

$$\hat{\theta}_{\lambda,f} = \arg\min_{\theta\in\mathbb{R}^n} \frac{1}{2}\|y - \theta\|^2 + \lambda f(\theta).$$

Here, we're using $\theta$ instead of $f$ since we're always considering fixed design, i.e., the data is just an $n$-dimensional vector. With the fact that we don't have any constraint on $\mathscr{F}$, we might just forget about functions at all. Let's see some examples of penalty functions.

**Example** ($\ell_1$-norm). If $\theta^*$ is sparse, then we can use the $\ell_1$-norm penalty $f(\cdot) = \|\cdot\|_1$.

**Example** (Nuclear-norm). If $\theta^*$ (a matrix) is low-rank, we can use the nuclear norm as $f$.

**Example** (Fused Lasso). If $f(\theta) = \sum_i |\theta_{i+1} - \theta_i| = \mathrm{TV}(\theta)$, then $\theta^*$ is piecewise constant.

**Example** (Trend Filtering). If $f(\theta) = \sum_i |\theta_{i+1} - 2\theta_i + \theta_i|$, then $\theta^*$ is piecewise linear.

The above are all $\ell_1$-type penalty. One can also consider $\ell_2$-type penalty.

[5] It's easy to check that $\mathrm{Star}(A, A) = \mathrm{conv}(A)$ for any set $A$.

**Example** (Ridge regression). $f(\theta) = \|\theta\|_2^2$.

There are some non-trivial differences between the penalized least square and constrained least square. Consider the following example.

**Example** (Constrained fused Lasso v.s. penalized fused Lasso). Let $f(\theta) = \mathrm{TV}(\theta)$, then the constrained least square estimator for $\mathscr{F}_v = \{\theta\colon \mathrm{TV}(\theta) \le v\}$ is

$$\hat{\theta}_v = \underset{\theta\colon \mathrm{TV}(\theta)\le v}{\arg\min} \|y-\theta\|^2,$$

while the penalized least square estimator is

$$\hat{\theta}_{\lambda,\mathrm{TV}} = \underset{\theta\in\mathbb{R}^n}{\arg\min} \|y-\theta\|^2 + \lambda\,\mathrm{TV}(\theta).$$

Observe that for any fixed $y$, $\hat{\theta}_{\lambda,\mathrm{TV}} = \hat{\theta}_v$ for some $v$ depends on $y$. That is to say, in general,

$$\|\hat{\theta}_v - \theta^*\|^2 \neq \|\hat{\theta}_{\lambda,\mathrm{TV}} - \theta^*\|^2.$$

**Remark.** Even though we have some well-established theory around constrained least square, the penalized version is more popular because of computational concerns.

Clearly, $\hat{\theta}_{\lambda,f}$ is an $M$-estimator, so we're going to use our established theory, i.e., the non-asymptotic rate of convergence. First, fix some $\lambda$ and $f$, we may write $\hat{\theta}_{\lambda,f} = \hat{\theta}$. Then, by plugging in $y = \theta^* + \epsilon$,

$$\hat{\theta} = \underset{\theta\in\mathbb{R}^n}{\arg\max} \left( \langle \epsilon, \theta - \theta^* \rangle - \frac{1}{2}\|\theta-\theta^*\|^2 - \lambda f(\theta) \right)$$

where we omit $\|\epsilon\|/2$ since it's independent of $\theta$. Hence, we may define[6]

$$M_n(\theta) \coloneqq \langle \epsilon, \theta - \theta^* \rangle - \frac{1}{2}\|\theta-\theta^*\|^2 - \lambda f(\theta).$$

Correspondingly, let

$$M(\theta) \coloneqq -\frac{1}{2}\|\theta-\theta^*\|^2$$

with

$$\theta^* = \underset{\theta\in\mathbb{R}^n}{\arg\max}\, M(\theta).$$

**Intuition.** Here, $M_n(\theta) \neq \mathbb{E}\left[M(\theta)\right]$, where we have additional terms. However, if the penalty is "appropriate", the actual penalization would be small. In other case, even if the penalty is not small, one may want to use it to control the "complexity" of the estimator.[a]

[a] This is the same thing as regularization in, e.g., machine learning.

We first check the following.

**Claim.** $d(\theta,\theta') = \|\theta-\theta'\|_2$ satisfies the growth condition.

**Proof.** Since $d^2(\theta,\theta') = \|\theta-\theta'\|^2 \ge -\frac{1}{2}\left(\|\theta-\theta^*\|^2 - \|\theta'-\theta^*\|^2\right) = M(\theta) - M(\theta')$. ⊛

Then, our next goal is to bound the localized empirical process

$$\sup_{\substack{\theta\in\mathbb{R}^n \\ \|\theta-\theta^*\|\le t}} (M_n - M)(\theta) - (M_n - M)(\theta^*) = \langle \epsilon, \theta-\theta^*\rangle - \lambda f(\theta) + \lambda f(\theta^*).$$

[6] Note that previously we're dealing with minimum, so here the signs are flipped.

### 5.4.1 Convex Penalty

To control the term $\lambda(f(\theta^*) - f(\theta))$, we assume that $f$ is convex. In this case, we can linearize $f$ by

$$f(\theta) \geq f(\theta^*) + \langle s, \theta - \theta^* \rangle,$$

i.e., $s$ is a sub-gradient.

**Definition 5.4.1** (Sub-gradient). The set of *sub-gradients* $\partial f(x^*)$ of $f : \mathbb{R}^n \to \mathbb{R}$ at $x^*$ contains $s \in \mathbb{R}^n$ such that for every $x \in \mathbb{R}^n$,

$$f(x) \geq f(x^*) + \langle s, x - x^* \rangle.$$

Hence, for convex $f$, we may take any $s \in \partial f(\theta^*)$ such that

$$\sup_{\substack{\theta \in \mathbb{R}^n \\ \|\theta - \theta^*\| \leq t}} \langle \epsilon, \theta - \theta^* \rangle - \lambda f(\theta) + \lambda f(\theta^*) \leq \sup_{\substack{\theta \in \mathbb{R}^n \\ \|\theta - \theta^*\| \leq t}} \langle \epsilon, \theta - \theta^* \rangle - \lambda \langle s, \theta - \theta^* \rangle = \sup_{\substack{\theta \in \mathbb{R}^n \\ \|\theta - \theta^*\| \leq t}} \langle \epsilon - \lambda s, \theta - \theta^* \rangle.$$

By Cauchy-Schwarz, this is further upper-bounded by $t\|\epsilon - \lambda s\|$. Since this is true for any $s \in \partial f(\theta^*)$,

$$\sup_{\substack{\theta \in \mathbb{R}^n \\ \|\theta - \theta^*\| \leq t}} (M_n - M)(\theta) - (M_n - M)(\theta^*) \leq t \cdot \inf_{s \in \partial f(\theta^*)} \|\epsilon - \lambda s\| = t \cdot \operatorname{dist}(\epsilon, \lambda \cdot \partial f(\theta^*)).$$

By taking the expectation, we successfully bound the localized empirical process by $\phi_n(t)$ where

$$\phi_n(t) \coloneqq t \cdot \mathbb{E}\left[\operatorname{dist}(\epsilon, \lambda \cdot \partial f(\theta^*))\right].$$

It's clear that $\phi_n(t)$ satisfies the sub-quadratic assumption with $\alpha = 1$ since $\phi_n(ct) = c^\alpha \phi_n(t)$ for $\alpha = 1$. As the last step, we consider the rate-determining equation $\phi_n(\delta_n) \approx \delta_n^2$, which gives

$$\delta_n \approx \mathbb{E}\left[\operatorname{dist}(\epsilon, \lambda \partial f(\theta^*))\right],$$

which implies $\|\hat{\theta} - \theta^*\| = O_p(\delta_n)$.

### 5.4.2 A Deterministic Approach

Without going through this machinery, one can directly show the deterministic bound indeed [OH13].

**Lemma 5.4.1** (Oymak-Hassibi). For a penalized least square problem, $\|\hat{\theta} - \theta^*\| \leq \operatorname{dist}(\epsilon, \lambda \partial f(\theta^*))$.

**Proof.** First, $\hat{\theta}$ minimizes $g$ if and only if $0 \in \partial g(\hat{\theta})$, where $g$ is defined as

$$g(\theta) = \frac{1}{2}\|y - \theta\|^2 + \lambda f(\theta).$$

Equivalently, $0 \in \hat{\theta} - y + \lambda \cdot \partial f(\hat{\theta})$, which implies $y - \hat{\theta} \in \lambda \cdot \partial f(\hat{\theta})$.

**Claim.** If $f : \mathbb{R}^n \to \mathbb{R}$ is convex, then for any $\theta_1, \theta_2 \in \mathbb{R}^n$, $s_1 \in \partial f(\theta_1)$, $s_2 \in \partial f(\theta_2)$,

$$\langle \theta_1 - \theta_2, s_1 - s_2 \rangle \geq 0.$$

**Proof.** Since $f(\theta_1) \geq f(\theta_2) + \langle s_2, \theta_1 - \theta_2 \rangle$ and $f(\theta_2) \geq f(\theta_1) + \langle s_1, \theta_2 - \theta_1 \rangle$, adding them together gives the result. ⊛

Hence, for any $s \in \partial f(\theta^*)$, $\langle y - \hat{\theta} - \lambda s, \hat{\theta} - \theta^* \rangle \geq 0$. By writing $y = \theta^* + \epsilon$, we finally get

$$\|\hat{\theta} - \theta^*\|^2 \leq \langle \epsilon - \lambda s, \hat{\theta} - \theta^* \rangle \leq \|\epsilon - \lambda s\| \|\hat{\theta} - \theta^*\|.$$

Since this holds for any $s \in \partial f(\theta^*)$, taking the infimum over $s \in \partial f(\theta^*)$ gives the result. ■

**Remark.** This bound is good if $\partial f(\theta^*)$ is large, i.e., this is useful when $f$ is not differentiable.

### 5.4.3 Sparse Means

Let's see an application. Let $y = \theta^* + \epsilon$ where $\theta^*$ is $k$-sparse, so it's natural to consider $f(\theta) = \|\theta\|_1$, i.e.,

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^n} \frac{1}{2}\|y - \theta\|^2 + \lambda\|\theta\|_1.$$

This problem has a closed-form solution, but let's see how can our technique help us. First, we see that

$$\partial f(\theta^*) = \left\{ v \in \mathbb{R}^n \colon v_i = \begin{cases} 1, & \text{if } \theta_i^* > 0; \\ -1, & \text{if } \theta_i^* < 0; \\ [-1, 1], & \text{if } \theta_i^* = 0. \end{cases} \right\}.$$

**Intuition.** We see that $\partial f(\theta^*)$ can be potentially large.

Then, from Lemma 5.4.1 (actually, the expectation version),

$$\|\hat{\theta} - \theta^*\|^2 \le \mathbb{E}\left[\operatorname{dist}^2(\epsilon, \lambda\partial f(\theta^*))\right] = \mathbb{E}\left[\sum_{i \colon \theta_i^* \neq 0} (\epsilon_i - \lambda \operatorname{sgn}(\theta_i^*))^2\right] + \mathbb{E}\left[\sum_{i \colon \theta_i^* = 0} \left(\epsilon_i - \operatorname{Proj}_{[-\lambda,\lambda]}(\epsilon_i)\right)^2\right],$$

where $\operatorname{Proj}_{[-\lambda,\lambda]}(\epsilon_i)$ is the projection of $\epsilon_i$ to the interval $[-\lambda, \lambda]$. Thus, for all $i$, let

$$\operatorname{Soft}_\lambda(\epsilon_i) \coloneqq \epsilon_i - \operatorname{Proj}_{[-\lambda,\lambda]}(\epsilon_i) = \begin{cases} \epsilon_i - \lambda, & \text{if } \epsilon_i > \lambda; \\ 0, & \text{if } \epsilon_i \in [-\lambda, \lambda]; \\ \epsilon_i + \lambda, & \text{if } \epsilon_i < -\lambda. \end{cases}.$$

**Note.** One can show that $\hat{\theta}_i = \operatorname{Soft}_\lambda(y_i)$.

With $k = \|\theta^*\|_0$, we therefore have

$$\|\hat{\theta} - \theta^*\|^2 \le \mathbb{E}\left[\operatorname{dist}^2(\epsilon, \partial f(\theta^*))\right] \le k(1 + \lambda^2) + (n - k) \cdot \mathbb{E}\left[\left(\operatorname{Soft}_\lambda(Z)\right)^2\right]$$

for $Z \sim \mathcal{N}(0, 1)$.

**Claim.** For $Z \sim \mathcal{N}(0, 1)$ and $\operatorname{Soft}_\lambda$ defined above,

$$\mathbb{E}\left[\left(\operatorname{Soft}_\lambda(Z)\right)^2\right] \le \frac{2e^{-\lambda^2/2}}{\lambda\sqrt{2\pi}}.$$

**Proof.** By a direct calculation, we have

$$\begin{aligned}
\mathbb{E}\left[(\operatorname{Soft}_\lambda(Z))^2\right] &= \int_{-\infty}^{\infty} (\operatorname{Soft}_\lambda(z))^2 \phi(z)\,\mathrm{d}z \\
&= 2\int_0^{\infty} (\operatorname{Soft}_\lambda(z))^2 \phi(z)\,\mathrm{d}z \\
&= 2\int_\lambda^{\infty} (z - \lambda)^2 \phi(z)\,\mathrm{d}z \\
&= 2\left[\int_\lambda^{\infty} z^2\phi(z)\,\mathrm{d}z - 2\lambda\int_\lambda^{\infty} z\phi(z)\,\mathrm{d}z + \lambda^2\int_\lambda^{\infty} \phi(z)\,\mathrm{d}z\right] \\
&= 2(1 + \lambda^2)(1 - \Phi(\lambda)) - 2\lambda\phi(\lambda).
\end{aligned}$$

From the Gaussian tail bound, $1-\Phi(\lambda) \le \phi(\lambda)/\lambda$ for any $\lambda > 0$, hence

$$\mathbb{E}\left[(\mathrm{Soft}_\lambda(Z))^2\right] \le 2(1+\lambda^2)\frac{\phi(\lambda)}{\lambda} - 2\lambda\phi(\lambda) = 2\frac{\phi(\lambda)}{\lambda} = \frac{2\exp(-\lambda^2/2)}{\lambda\sqrt{2\pi}}.$$

⊛

Hence, our final bound becomes

$$\|\hat{\theta}-\theta^*\|^2 \le k(1+\lambda^2) + (n-k)\sqrt{\frac{2}{\pi}}\frac{e^{-\lambda^2/2}}{\lambda}$$

with $\lambda = \sqrt{2\log n/k}$,

$$= k\left(1+2\log\frac{n}{k}\right) + (n-k)\sqrt{\frac{2}{\pi}}\frac{k}{n}\sqrt{\frac{1}{2\log n/k}} = 2k\log\frac{n}{k}(1+o(1))$$

w.r.t. $k/n \to 0$. Finally, we note that

$$\mathbb{E}\left[\frac{1}{n}\|\hat{\theta}-\theta^*\|^2\right] \le 2\frac{k}{n}\log\frac{n}{k}(1+o(1)).$$

**Note.** In general, if $\theta^*$ is $k$-sparse, $\ell_1$ penalization with the right penalty can obtain a $\frac{k}{n}\log\frac{n}{k}$ rate.

**Remark.** $\lambda = \sqrt{2\log n/k}$ is optimal, even regarding the 2 factor in the final bound $2k\log n/k$.

But setting $\lambda$ requires some knowledge of $k$, hence the structure of $\theta^*$. In this is not the case, we just set $\lambda = \sqrt{2\log n}$, yielding a bound of $2k\log n(1+o(1))$. If $k$ is large, then this bound is not tight.

**Note.** This technique of bounding distant to the sub-gradient set has been applied to show fast rates in other settings, such as nuclear-norm penalization [OH10], or TV penalty ($\theta^*$ is $k$-sparse), and filtering.

# Lecture 30: Union of Subspaces Estimator

## 5.4.4 Union of Subspaces Theorem

13 Nov. 9:00

As our last section about fixed design non-parametric regression, we're going to look at one more example of penalized least square. Specifically, we consider the setup where $\theta$ comes from a union of subspaces of $\mathbb{R}^n$, and penalize the solution $\theta$ by the dimension of the subspace it lies in.

**Intuition.** In fact, it's like a combination of constrained least square and penalized least square.

We first need a lemma before stating the general result in this section.

**Lemma 5.4.2.** For any subspace $S \subseteq \mathbb{R}^n$ with $Z \sim \mathcal{N}(0, \sigma^2 I_n)$, with probability $\ge 1-e^{-t/2}$,

$$\sup_{\theta\in S}\left\langle Z, \frac{\theta-\theta^*}{\|\theta-\theta^*\|}\right\rangle^2 \le 2\dim(S) + 4(1+t).$$

**Proof.** Consider two cases.

- $\theta^* \in S$: The supremum becomes $\sup_{\theta\in S}\langle Z, \frac{\theta}{\|\theta\|}\rangle^2 = \sup_{\theta\in S}|\operatorname{Proj}_S Z|^2$, where $|\operatorname{Proj}_S Z|^2 \sim \chi^2_{\dim(S)}$, which follows a SubExp-tail. By the standard concentration result, we're done.

- $\theta^* \notin S$: Denote $\text{Proj}_S(\theta) =: P_S(\theta)$, then

$$\left\langle Z, \frac{\theta - \theta^*}{\|\theta - \theta^*\|} \right\rangle = \left\langle Z, \frac{\theta - P_S\theta^* - (I - P_S)\theta^*}{\|\theta - P_S\theta^* - (I - P_S)\theta^*\|} \right\rangle$$
$$= \underbrace{\left\langle Z, \frac{\theta - P_S\theta^*}{\|\theta - P_S\theta^* - (I - P_S)\theta^*\|} \right\rangle}_{T_1} + \underbrace{\left\langle Z, \frac{(I - P_S)\theta^*}{\|\theta - P_S\theta^* - (I - P_S)\theta^*\|} \right\rangle}_{T_2},$$

which implies $\langle Z, \frac{\theta-\theta^*}{\|\theta-\theta^*\|}\rangle^2 = T_2^2 + T_2^2$.[a] Hence, it suffices to bound $\sup_{\theta\in S} T_1$ and $\sup_{\theta\in S} T_2$, respectively. For $T_1$, we have

$$\sup_{\theta\in S} T_1 = \sup_{\theta\in S} \left\langle Z, \frac{\theta - P_S\theta^*}{\|\theta - P_S\theta^* - (I - P_S)\theta^*\|} \right\rangle \le \sup_{\theta\in S:\ \|\theta\|\le 1} \langle Z, \theta\rangle = \|P_S Z\| \ge \sqrt{\dim(S)} + t$$

with probability $\le Ce^{-Ct}$. For $T_2$, let $S'$ denote the 1-dimensional subspace spanned by $(I - P_S)\theta^*$, then

$$\sup_{\theta\in S} T_2 = \sup_{\theta\in S':\ \|\theta\|\le 1} \langle Z, \theta\rangle = \|P_{S'} Z\| > 1 + t$$

with probability $\le Ce^{-Ct}$. Combining these two upper-bounds, we're done.

■

[a] Since if $a \in S$, $b \in S^\perp$, we have $\|a + b\|^2 = \|a\|^2 + \|b\|^2$ and $\|a + b\| \ge \|a\|$.

Now, let $y = \theta^* + \epsilon$ where $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, i.e., the sequence model. Then we have the following.

**Theorem 5.4.1.** Let $\mathcal{S}$ be a finite collection of subspaces of $\mathbb{R}^n$ and $\theta^* \in \bigcup_{S\in\mathcal{S}} S$. Define

$$\hat{\theta} = \underset{\theta\in\bigcup_{S\in\mathcal{S}} S}{\arg\min} \|y - \theta\|^2 + \lambda k(\theta)$$

where $k(\theta) = \min\{\dim(S) \colon \theta \in S, S \in \mathcal{S}\}$. If $|\{S \in \mathcal{S} \colon \dim(S) = k\}| \le n^{ck}$ for all $k = 0, , \ldots, n$ and $\lambda > c\sigma^2 \log n$ for some constant $c$, then with high probability,

$$\frac{1}{n}\|\hat{\theta} - \theta^*\|^2 \le \frac{c\lambda k(\theta^*)}{n}.$$

**Proof.** Since $\|y - \hat{\theta}\|^2 + \lambda k(\hat{\theta}) \le \|y - \theta^*\|^2 + \lambda k(\theta^*)$, we have

$$\|\hat{\theta} - \theta^*\|^2 \le 2\langle \sigma Z, \hat{\theta} - \theta^*\rangle - \lambda k(\hat{\theta}) + \lambda k(\theta^*)$$

assume $\sigma^2 = 1$ for simplicity,

$$\le 2\left\langle Z, \frac{\hat{\theta} - \theta^*}{\|\hat{\theta} - \theta^*\|} \right\rangle \cdot \|\hat{\theta} - \theta^*\| - \lambda k(\hat{\theta}) + \lambda k(\theta^*)$$
$$\le 2\left\langle Z, \frac{\hat{\theta} - \theta^*}{\|\hat{\theta} - \theta^*\|} \right\rangle^2 + \frac{1}{2}\|\hat{\theta} - \theta^*\|^2 - \lambda k(\hat{\theta}) + \lambda k(\theta^*)$$

since $2ab \le 2a^2 + b^2/2$ where $a = \langle Z, \frac{\hat{\theta}-\theta^*}{\|\hat{\theta}-\theta^*\|}\rangle$ and $b = \|\hat{\theta} - \theta^*\|$.[a] Hence,

$$\frac{1}{2}\|\hat{\theta} - \theta^*\|^2 \le \underbrace{2\left\langle Z, \frac{\hat{\theta} - \theta^*}{\|\hat{\theta} - \theta^*\|} \right\rangle^2}_{R} - \lambda k(\hat{\theta}) + \lambda k(\theta^*).$$

We note that $\lambda k(\theta^*)$ is a good term, so our goal is to bound $R$ now. Specifically,

$$\begin{aligned} R &\le \max_{1\le k\le n}\left\{\max_{S\in\mathcal{S}:\ \dim(S)=k}\left\{\sup_{\theta\in S}\left(2\left\langle Z, \frac{\theta-\theta^*}{\|\theta-\theta^*\|}\right\rangle^2 - \lambda k(\theta)\right)\right\}\right\} \\ &= \max_{1\le k\le n}\underbrace{\left\{\max_{S\in\mathcal{S}:\ \dim(S)=k}\left\{\sup_{\theta\in S}\left(2\left\langle Z, \frac{\theta-\theta^*}{\|\theta-\theta^*\|}\right\rangle^2 - \lambda k\right)\right\}\right\}}_{E_k} \end{aligned}$$

since in the supremum, $\theta\in S$ with $\dim(S)=k$, $k(\theta)=k$. From Lemma 5.4.2,[b]

$$\mathbb{P}\left(\sup_{\theta\in S}\left(\left\langle Z, \frac{\theta-\theta^*}{\|\theta-\theta^*\|}\right\rangle^2 - \lambda k\right) > t\right) \lesssim e^{-(\lambda-2)k}e^{-t},$$

with union bounds over $\{S\in\mathcal{S}:\ \dim(S)=k\}$ (with size $\le n^{ck}$) gives

$$\mathbb{P}(E_k > t) \lesssim n^{ck}e^{-\lambda k}e^{-t} = e^{ck\log n-\lambda k-t} \le e^{-ck\log n-t}.$$

By choosing $\lambda > 2c\log n$, we further have

$$\mathbb{P}(R>t) = \mathbb{P}\left(\max_{1\le k\le n}E_k > t\right) \le \sum_{k=1}^{n} e^{-ck\log n}e^{-t} \le ce^{-t}.$$

In fact, the inequality holds as long as $\lambda > c\log n$. Consequently,

$$\mathbb{P}\left(\frac{1}{2}\|\hat\theta-\theta^*\|^2 > \lambda k(\theta^*)+t\right) \le ce^{-t}.$$

Finally, by treating $c$ as a universal constant that may vary yields the result. ■

[a] Here, we do not use $2ab\le a^2+b^2$ since we really want $c\|\hat\theta-\theta^*\|^2$ for some $c<1$, so we can proceed.
[b] Note that $2\langle Z, \frac{\theta-\theta^*}{\|\theta-\theta^*\|}\rangle^2 - \lambda k > t \Leftrightarrow 2\langle Z, \frac{\theta-\theta^*}{\|\theta-\theta^*\|}\rangle^2 > t+\lambda k \Leftrightarrow 2\langle Z, \frac{\theta-\theta^*}{\|\theta-\theta^*\|}\rangle^2 - 2k > t+(\lambda-2)k$.

**Notation** (Oracle rate). The $c\lambda k(\theta^*)/n$ in the result of Theorem 5.4.1 is called the *oracle rate*.

**Remark.** If $S\ni\theta^*$ is known, then we just do a projection on the subspace $S$, which gives a bound on the mean square error as $\le k(\theta^*)/n$.

**Intuition.** Hence, $\lambda\approx\log n$ is like a "search cost" we need to pay.

**Example** (High dimensional linear regression). Let $X\in\mathbb{R}^{n\times p}$ to be the design matrix with $p>n$, and the subspaces are indexed by $S\subseteq[p]$. For each $S$, the corresponding subspace is

$$\mathcal{C}(X_S) = \{X_S\beta:\ \beta\in\mathbb{R}^{|S|}\},$$

i.e., the column space. Assume that $\bigcup_{S\subseteq[p]}\mathcal{C}(X_S)=\mathcal{C}(X)=\mathbb{R}^n$, then $k(\theta)=\min\{\|\beta\|_0:\ X\beta=\theta\}$. Denote $\theta^*=X\beta^*$, the corresponding estimator is $\hat\theta=X\hat\beta$, where

$$\hat\beta = \arg\min_{\beta\in\mathbb{R}^p}\|y-X\beta\|^2+\lambda\|\beta\|_0,$$

i.e., the *BIC estimator*. In this case, without any assumptions on $X$, the oracle rate is

$$\frac{1}{n}\|X\hat\beta-X\beta^*\|^2 \le c\sigma^2\|\beta^*\|_0\frac{\log n\log p}{n}.$$

**Proof.** To use Theorem 5.4.1, we first bound $|\{S \in \mathcal{S} \colon \dim(S) = k\}|$. We have

$$\#\text{subspaces with dimension } k \le \binom{p}{k} = n^{\log \binom{p}{k}} \le n^{ck \log p}.$$

Hence, by choosing $\lambda > c\sigma^2 \log n$, with $\dim(\beta^*) \le \|\beta^*\|_0$, we have the result. ⊛

**Example** (Piecewise costant 1D signal). Let $\pi$ be an interval partition of $[n]$, and let

$$S_\pi = \{\theta \in \mathbb{R}^n \colon \theta \text{ piecewise constant on } \pi\},$$

then $\dim(S_\pi) = \#\text{blocks of } \pi =: |\pi|$. Hence, $k(\theta) = \min\{|\pi| \colon \theta \in S_\pi\} = \#\text{pieces of } \theta$. Then

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^n} \|y - \theta\|^2 + \lambda k(\theta) = \arg\min_{\pi \in \mathcal{P}} \|y - \operatorname{Proj}_\pi y\|^2 + \lambda |\pi|,$$

where $\mathcal{P}$ is the set of all interval partitions of $[n]$. This is a discrete optimization problem and can be computed by dynamic programming in $O(n^3)$ time. In particular, the oracle rate is

$$\frac{1}{n}\|\hat{\theta} - \theta^*\|^2 \le \frac{ck(\theta^*)}{n}\sigma^2 \log n.$$

**Proof.** By choosing $\lambda > c\sigma^2 \log n$, $\hat{\theta}$, from Theorem 5.4.1 with

$$|\{S \colon \dim(S) = k\}| = |\{\pi \in \mathcal{P} \colon |\pi| = k\}| \le |\{\pi \in \mathcal{P} \colon |\pi| \le k\}| = \binom{n+k-1}{k-1} \le n^{ck}$$

for some constant $c$,[a] we're done. ⊛

[a] It's just the number of ways of putting $k-1$ bars between $n$ balls.

**Example** (Piecewise constant 2D signal on rectangle; decision tree). Consider decision trees, which can be viewed as a rectangle partition of $[n] \times [n]$, partitioned hierarchically. Then, consider the set of all decision trees $\mathcal{D}$, and for each rectangle partition $\pi \in \mathcal{D}$, let

$$S_\pi = \{\theta \in \mathbb{R}^{n \times n} \colon \theta \text{ is constant on every rectangle in } \pi\},$$

hence $k(\theta) = \min\{|\pi| \colon \pi \in \mathcal{D}, \theta \text{ is constant on every rectangle of } \pi\}$. Correspondingly, let

$$\hat{\theta} = \arg\min_{\pi \in \mathcal{D}} \|y - \operatorname{Proj}_\pi y\|^2 + \lambda |\pi|$$

where $|\pi|$ is the number of cells of the corresponding decision tree. Surprisingly, $\hat{\theta}$ can be computed in $O(n^5)$ time using dynamic programming. In particular, the oracle rate is again

$$\frac{1}{n}\|\hat{\theta} - \theta^*\|^2 \le \frac{ck(\theta^*)}{n}\sigma^2 \log n.$$

**Proof.** The number of subspaces with dimension $k$ is upper-bounded by $\binom{n}{k} \le n^{4k}$, hence Theorem 5.4.1 gives the desired result. ⊛

**Note.** The above two results are useful when $\theta^*$ is piecewise constant, i.e., $k(\theta^*)$ is small. This is usually the case for structured signals, e.g., images for the 2D case.

# Chapter 6

# Epilogue

## Lecture 31: Rademacher Complexity for Neural Networks

At the end of the course, we're going to discuss some caveats of the theory we have developed and conclude with some modern applications to neural networks. 14 Nov. 9:00

## 6.1 Large Margin Theory for Classification

### 6.1.1 Classification in Practice

Throughout the course, when talking about bounding the excess risk for empirical risk minimization for classification,[1] we consider the expected supremum of empirical process, which in turns depends on either $\mathrm{VC}(\mathscr{F})$ or the Rademacher complexity $R(\mathscr{F})$ for some boolean function classes $\mathscr{F}$.

However, in practice, we consider real-valued function class $\mathscr{F}$, and consider either the VC dimension or the Rademacher complexity of $\operatorname{sgn}(\mathscr{F})$, making which boolean. The tricky part here is that it's very likely to have

$$R(\mathscr{F}) < R(\operatorname{sgn}(\mathscr{F})),$$

which makes our developed bounds potentially loose.

**Example.** For $\mathscr{F} = \{x^\top \beta \colon \|\beta\|_2 \le 1, \|x\|_2 \le 1\}$. In this case, we have $R_n(\mathscr{F}) \le c/\sqrt{n}$ times some dimension-free rate using fat-shattering argument (Theorem 3.3.7). On the other hand, we have

$$R_n(\operatorname{sgn}(\mathscr{F})) \approx \sqrt{d/n} \approx \mathrm{VC}(\operatorname{sgn}(\mathscr{F})).$$

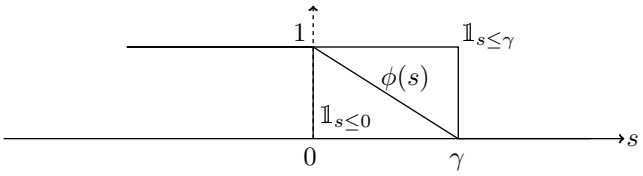
### 6.1.2 Large Margin Theorem

The goal is to get a bound on the excess risk for the binary classification problem in terms of $R_n(\mathscr{F})$, and not on $R_n(\operatorname{sgn}(\mathscr{F}))$. We start by fixing a margin parameter $\gamma > 0$, and define

$$\phi(s) = \begin{cases} 1, & \text{if } s \le 0; \\ 1 - s/\gamma, & \text{if } 0 < s < \gamma; \\ 0, & \text{if } s \ge \gamma. \end{cases}$$

Assume that we're considering a binary classification problem with labels being $y \in \{\pm 1\}$ under the empirical risk minimization setting. Then,

$$\mathbb{1}_{yf(x)\le 0} \le \phi(yf(x)) \le \mathbb{1}_{yf(x)\le\gamma}.$$




[1] Recall the 1D classification example.

By taking the expectation, we have

$$\mathbb{E}\left[\mathbb{1}_{yf(x)\le 0}\right] - \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{y_i f(x_i)\le\gamma} \le \mathbb{E}\left[\phi(yf(x))\right] - \frac{1}{n}\sum_{i=1}^{n}\phi(y_i f(x_i))$$
$$\le \sup_{f\in\mathscr{F}}\left[\mathbb{E}\left[\phi(yf(x))\right] - \frac{1}{n}\sum_{i=1}^{n}\phi(y_i f(x_i))\right]$$

i.e., a supremum of an empirical process, and by the McDiarmid inequality,

$$\le \mathbb{E}\left[\sup_{f\in\mathscr{F}}\left[\mathbb{E}\left[\phi(yf(x))\right] - \frac{1}{n}\sum_{i=1}^{n}\phi(y_i f(x_i))\right]\right] + \frac{u}{\sqrt{n}}$$

with probability at least $1-e^{-2u^2}$ since $\phi$ is bounded between 0 and 1, which is further bounded by

$$\le 2\mathbb{E}\left[\sup_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\phi(y_i f(x_i))\right] + \frac{u}{\sqrt{n}} \qquad \text{symmetrization}$$
$$\le \frac{2}{\gamma}\mathbb{E}\left[\sup_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(y_i f(x_i))\right] + \frac{u}{\sqrt{n}} \qquad \text{contraction principle}$$

since $\phi$ is $\frac{1}{\gamma}$-Lipschitz, and finally,

$$= \frac{2}{\gamma}\underbrace{\mathbb{E}\left[\sup_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(x_i)\right]}_{R_n(\mathscr{F})} + \frac{u}{\sqrt{n}}$$

since $y_i\epsilon_i$ is no difference from $\epsilon_i$. After rearranging, we get the following.

**Theorem 6.1.1** (Large margin theorem). Given any real function class $\mathscr{F}$, for any predictor $f\in\mathscr{F}$ of a binary classification problem with label being $\{\pm 1\}$ in the empirical risk minimization setting, with probability at least $1-e^{2u^2}$,

$$\mathbb{E}\left[\mathbb{1}_{yf(x)\le 0}\right] \le \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{y_i f(x_i)\le\gamma} + \frac{2}{\gamma}R_n(\mathscr{F}) + \frac{u}{\sqrt{n}}.$$

**Notation** (Margin error). The training *margin error* of a predictor $f$ is given by $\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{y_i f(x_i)\le\gamma}$.

We see that the left-hand side is the test error of $\operatorname{sgn}(\mathscr{F})$, while the right-hand side is the training margin error plus the Rademacher complexity.

**Note.** There are a couple of messages from the large margin theorem:

(a) The test error is in terms of $R_n(\mathscr{F})$, not $R_n(\operatorname{sgn}(\mathscr{F}))$.

(b) We paid a cost, i.e., the margin error, instead of the usual training error, which is larger.

(c) We need a version that holds for a grid of $\gamma$, which can be done by a union-bound argument.

(d) It gives one explanation of why test error can get better even if the training error is 0, i.e., the predictor generalizes.

**Intuition.** The above suggests that we should minimize the sum between the margin error and the Rademacher complexity, which is usually better from the usual bound.[a]

[a] I.e., the sum between a smaller training error plus an often larger $R_n(\operatorname{sgn}(\mathscr{F}))$.

## 6.2 Rademacher Complexity for Neural Networks

Recall the following.

**As previously seen.** For linear functions, Theorem 3.3.7 tells us that norm constraints can help us to get a better bound on the Rademacher complexity.

Such a bound is much less than the number of parameters, i.e., the dimension, and in fact, similar things happen for other function classes. We will see this for (classical) neural networks.

### 6.2.1 Multi-Layer Perceptron

Consider the following classical neural network called multi-layer perceptron.

**Definition 6.2.1** (Multi-layer perceptron). Let $\sigma\colon \mathbb{R} \to \mathbb{R}$ to be a non-linear activation function that is applied coordinate-wise, and let the weight be $\theta = (W_1, \dots, W_L)$ for $W_i\colon \mathbb{R}^{d_{i-1}} \to \mathbb{R}^{d_i}$ [a] with $d_0 \coloneqq d$ and $d_L \coloneqq 1$. Then, the *multi-layer perceptron* is defined as $f\colon \mathbb{R}^d \to \mathbb{R}$ such that for $x \in \mathbb{R}^d$,

$$f_\theta(x) \coloneqq W_L\sigma(\dots\sigma(W_2\sigma(W_1x))).$$

[a] Here, $W_i$ can be thought as matrix in $\mathbb{R}^{d_i \times d_{i-1}}$, and the application is just matrix multiplication.

**Note.** One might instead define $f_\theta(x)$ by $\sigma(W_L\sigma(\dots\sigma(W_2\sigma(W_1x))))$, i.e., we apply an activation function at the end again.

Note that for the activation function $\sigma$, we usually want it to be Lipschitz and positive homogeneity ($\sigma(\alpha x) = \alpha\sigma(x)$ for $\alpha \geq 0$).

**Note.** If $\sigma$ is positive homogeneity, then it preserves the 0, i.e., $\sigma(0) = 0$.

A popular choice of $\sigma$ in machine learning is the following.

**Example** (ReLU). The ReLU activation function is defined as $\mathrm{ReLU}(x) = [x]_+ = \max(x, 0)$, which is 1-Lipschitz, positive homogeneity.

What we want is to look at the class of functions of neural networks with some complexity measure, say $\mathrm{comp}(\cdot)$, i.e.,

$$\mathscr{F} = \{f_\theta\colon \mathrm{comp}(\theta) \leq B\}.$$

The tricky part here is that there are so many ways to define complexity, and here, we're going to focus on norm-based constraints as we mentioned.

### 6.2.2 Norm-Based Constraint

Note that there are lots of invariants for multi-layer perceptron.

**Example** (Scaling). If we scale one layer by some factor $D > 0$, and divide the other layer by $D$, $f_\theta$ is unchanged if $\sigma$ is positive homogeneity.

Hence, for example, if we consider the Frobenius norm as the complexity measure, i.e.,

$$\mathrm{comp}(\theta) \coloneqq \sum_{i=1}^{L} \|W_i\|_F,$$

this won't capture the above scaling invariant. One potential fix could be using the product, i.e.,

$$\mathrm{comp}(\theta) \coloneqq \prod_{i=1}^{L} \|w_i\|_F,$$

which now respects the scaling invariant. But again, this may not respect some other notion of invariants. To see how to properly define our norm constraints, taking a recursive point of view when defining the multi-layer perceptron is helpful. Specifically, consider the following recursive definition.

**Definition 6.2.2** (Recursive construction of $\mathscr{F}_i$)**.** Given a base function class $\mathscr{F}_1 = \{f : \chi \to \mathbb{R}\}$ such that $0 \in \mathscr{F}_1$, the *recursive construction of* $\mathscr{F}_i$ is that for all $i > 1$, we define

$$\mathscr{F}_i \coloneqq \left\{ \sum_{j=1}^{d_{i-1}} w_j \sigma(f_j(x)) : f_j \in \mathscr{F}_{i-1}, w \in \mathbb{R}^{d_{i-1}} \right\}.$$

**Note.** It's convenient to write the weighted sum as an inner product, i.e.,

$$\sum_{j=1}^{d_{i-1}} w_j \sigma(f_j(x)) =: \langle w, (\sigma(f(x)))_j \rangle$$

where $(\sigma(f(x)))_j = (\sigma(f_1(x)), \sigma(f_2(x)), \ldots, \sigma(f_{d_{i-1}}(x)))^\top \in \mathbb{R}^{d_{i-1}}$.

**Example** (Multi-layer perceptron)**.** Multi-layer perceptron can be recursively constructed with $\mathscr{F}_1$ being linear.

We see that by viewing $w$ as the rows of some matrix $W \in \mathbb{R}^{d_i \times d_{i-1}}$ when constructing later layers, we recover multi-layer perceptron. This suggests we put norm constraints on these $w$. Turns out that in this case, we can consider the following norm, $\|\cdot\|_{p,q}$, defined for a matrix $A$, and by controlling this norm, we can obtain some non-trivial bound on $\mathscr{F}_L$.

**Definition 6.2.3** ($\|\cdot\|_{p,q}$)**.** For any matrix $A$, the norm $\|A\|_{p,q}$ is defined as

(a) first take the $\ell_p$ norm of all columns of $A$,

(b) and take $\ell_q$ norm of the $\ell_p$ norms we obtained for each column.

**Intuition.** Since $w$ can be thought of as the rows of some weight matrix $W$, controlling the norm of which can be done by $\|\cdot\|_{p,q}$ (with transpose).

Following the intuition, we put norm restriction on $w \in \mathbb{R}^{d_{i-1}}$ in our recursive construction of $\mathscr{F}_i$ for some $B_i \in \mathbb{R}$.

**Definition 6.2.4** (Recursive construction of $\mathscr{F}_i$ with norm constraint)**.** The *recursive construction of* $\mathscr{F}_i$ *with norm constraint* $B_i \in \mathbb{R}$ starts from a norm constrained base function class $\mathscr{F}_1 = \{f_w : \chi \to \mathbb{R} : \|w\|_1 \le B_1\}$[a] with $0 \in \mathscr{F}_1$, and for all $i > 1$,

$$\mathscr{F}_i \coloneqq \left\{ \sum_{j=1}^{d_{i-1}} w_j \sigma(f_j(x)) : f_j \in \mathscr{F}_{i-1}, w \in \mathbb{R}^{d_{i-1}}, \|w\|_1 \le B_i \right\}.$$

[a] For this definition to make sense, $f$ is now parametrized by some weights vector $w$.

With this construction, we can then define the norm-constrained multi-layer perceptron, starting from a linear function class as above. However, in this case, we should start from a norm-constrained linear function class instead.

**Definition 6.2.5** (Norm-constrained multi-layer perceptron)**.** The *norm-constrained multi-layer perceptron* $f_\theta(x) : \mathbb{R}^p \to \mathbb{R}$ is a function in $\mathscr{F}_L$ which is recursively constructed with norm constraint with the base function class being the norm-constrained linear function class

$$\mathscr{F}_1 = \{w^\top x : w \in \mathbb{R}^d, \|w\|_1 \le B_1\}.$$

With such recursive construction and the norm constraints, we can then bound the Rademacher complexity recursively as well.

**Intuition.** The general idea is that for the base function class $\mathscr{F}_1$, assume that $\|x_i\|_\infty \le 1$, then

$$\begin{aligned} R(\mathscr{F}_1) &= \mathbb{E}\left[\sup_{\|w\|_1 \le B_1} \frac{1}{n}\sum_{i=1}^n \epsilon_i \langle w, x_i\rangle\right] \\ &= \mathbb{E}\left[\sup_{\|w\|_1 \le B_1} \frac{1}{n}\left\langle w, \sum_{i=1}^n \epsilon_i x_i\right\rangle\right] \le \frac{1}{n}\sup_{\|w\|_1 \le B_1}\|w\|_1 \cdot \mathbb{E}\left[\left\|\sum_{i=1}^n \epsilon_i x_i\right\|_\infty\right] \le B_1\sqrt{\frac{\log d}{n}}. \end{aligned}$$

Consider the following.

**Lemma 6.2.1.** If $\mathscr{F}_i$ is defined recursively with norm constraints $B_i$ from the base function class $\mathscr{F}_1$ containing 0, and if $\sigma$ is 1-Lipschitz with $\sigma(0) = 0$, then for $x_1, \dots, x_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$,

$$R(\mathscr{F}_i) \le 2B_i R(\mathscr{F}_{i-1}).$$

**Proof.** From the definition of Rademacher complexity, we have

$$\begin{aligned} R(\mathscr{F}_i) &= \mathbb{E}\left[\sup_{\substack{\|w\|_1 \le B_i \\ f_j \in \mathscr{F}_{i-1}}} \frac{1}{n}\sum_{\ell=1}^n \epsilon_\ell \left(\sum_{j=1}^{d_{i-1}} w_j \sigma(f_j(x_\ell))\right)\right] \\ &= \mathbb{E}\left[\sup_{\substack{\|w\|_1 \le B_i \\ f_j \in \mathscr{F}_{i-1}}} \frac{1}{n}\sum_{j=1}^{d_{i-1}} w_j \cdot \sum_{\ell=1}^n \epsilon_\ell\left(\sigma(f_j(x_\ell))\right)\right] \\ &\le \sup_{\|w\|_1 \le B_i}\|w\|_1 \cdot \mathbb{E}\left[\sup_{f_j \in \mathscr{F}_{i-1}} \frac{1}{n}\left|\sum_{\ell=1}^n \epsilon_\ell \sigma(f_j(x_\ell))\right|\right] \le B_i \cdot \mathbb{E}\left[\sup_{f \in \mathscr{F}_{i-1}} \frac{1}{n}\left|\sum_{\ell=1}^n \epsilon_\ell \sigma(f(x_\ell))\right|\right]. \end{aligned}$$

Now, since

$$\sup_{f \in \mathscr{F}_{i-1}}\left|\sum_{\ell=1}^n \epsilon_\ell \sigma(f(x_\ell))\right| = \max\left(\sup_{f \in \mathscr{F}_{i-1}}\sum_{\ell=1}^n \epsilon_\ell \sigma(f(x_\ell)), \sup_{f \in \mathscr{F}_{i-1}} -\sum_{\ell=1}^n \epsilon_\ell \sigma(f(x_\ell))\right),$$

with the fact that $0 \in \mathscr{F}_1$ hence $0 \in \mathscr{F}_i$, we conclude that the above two terms are both $\ge 0$,

$$\mathbb{E}\left[\sup_{f \in \mathscr{F}_{i-1}} \frac{1}{n}\left|\sum_{\ell=1}^n \epsilon_\ell \sigma(f(x_\ell))\right|\right] \le \mathbb{E}\left[\sup_{f \in \mathscr{F}_{i-1}} \frac{1}{n}\sum_{\ell=1}^n \epsilon_\ell \sigma(f(x_\ell))\right] + \mathbb{E}\left[-\sup_{f \in \mathscr{F}_{i-1}} \frac{1}{n}\sum_{\ell=1}^n \epsilon_\ell \sigma(f(x_\ell))\right].$$

Since $-1$ and $\epsilon_\ell$ can be absorbed, we finally get

$$R(\mathscr{F}_i) \le 2B_i \cdot \mathbb{E}\left[\sup_{f \in \mathscr{F}_{i-1}} \frac{1}{n}\sum_{\ell=1}^n \epsilon_\ell \sigma(f(x_\ell))\right] = 2B_i R(\sigma \circ \mathscr{F}_{i-1}) \le 2B_i R(\mathscr{F}_{i-1}),$$

due to the contraction principle and the fact that $\sigma$ is 1-Lipschitz. ∎

Now, by using $\|\cdot\|_{1,\infty}$ as suggested above, we can obtain the following.

**Corollary 6.2.1.** If $\mathscr{F}_L$ is the class of norm-constrained multi-layer perceptrons with $\|W_i^\top\|_{1,\infty} \le B_i$,

$$R(\mathscr{F}_L) \le 2^L \prod_{i=1}^L B_i \sqrt{\frac{\log d}{n}}.$$

**Proof.** By iteratively applying Lemma 6.2.1 with the result from the intuition as a base case. ∎

This exponential dependency on the number of layers might not always be tight.

**Example** (Ultra-thin network). Consider an ultra-thin multi-layer perceptron with $\sigma = \text{ReLU}$ defined as

$$f(x) = \sigma(w_L \dots \sigma(w_2 \sigma(w_1^\top x)))$$

where $w_1 \in \mathbb{R}^d$, and $w_i \in \mathbb{R}$ for $i \geq 1$.[a] It's easy to see that the class of ultra-thin multi-layer perceptron is just the class of linear predictor $f(x) = \sigma(w^\top x)$ for some $w \in \mathbb{R}^d$. If we would like to put norm constraints on the latter, then we have

$$\mathscr{F}_{\text{linear}} = \{\sigma(w^\top x) \colon \|w\| \leq B\};$$

and equivalently, for the former,

$$\mathscr{F}_{\text{ultra-thin}} = \left\{ \sigma(w_L \dots \sigma(w_2 \sigma(w_1^\top x))) \colon \|w_1\| \cdot \prod_{j>1} |w_j| \leq M \right\}.$$

Now it's easy to see that no matter how large $L$ is, there will be no $2^L$ in $R(\mathscr{F}_{\text{linear}}) = R(\mathscr{F}_{\text{ultra-thin}})$ since the former is independent of $L$ in the first place.

[a] So everything is a scalar after the first layer, making $f$ ultra-thin (width 1).

A more sophisticated bound can be achieved by using the Frobenius norm.

**Theorem 6.2.1** ([GRS19]). Let $\mathscr{F}$ be the class of multi-layer perceptron with positive homogenous activation function $\sigma$ and $\|W_i\|_F \leq B_i$, then

$$R(\mathscr{F}) \leq \sqrt{\frac{L}{n}} \prod_{i=1}^{L} B_i \leq \frac{1}{n^{1/4}} \prod_{i=1}^{L} B_i.$$

# Appendix

# Appendix A

# Missing Proofs

In this chapter, we provide some missing proofs we omit in lectures, but solved in homework.

## A.1 Concentration Inequalities

### A.1.1 Hoeffding's Inequality

We start by providing the equivalent conditions of a random variable being sub-Gaussian.

**Lemma A.1.1** (Equivalent conditions (Lemma 2.3.1)). Given a random variable $X$ with $\mathbb{E}[X] = 0$, the following are equivalent for absolute constants $c_1, \dots, c_5 > 0$.

(a) $\mathbb{E}\left[e^{\lambda X}\right] \le e^{c_1^2 \lambda^2}$ for all $\lambda \in \mathbb{R}$.

(b) $\mathbb{P}(|X| \ge t) \le 2e^{-t^2/c_2^2}$.

(c) $(\mathbb{E}[|X|^p])^{1/p} \le c_3\sqrt{p}$.

(d) For all $\lambda$ such that $|\lambda| \le 1/c_4$, $\mathbb{E}\left[e^{\lambda^2 X^2}\right] \le e^{c_4^2\lambda^2}$.

(e) For some $c_5 < \infty$, $\mathbb{E}\left[e^{X^2/c_5^2}\right] \le 2$.

**Proof.** We show that (a) $\Rightarrow$ (b) $\Rightarrow$ (c) $\Rightarrow$ (d) $\Rightarrow$ (e) $\Rightarrow$ (a).

- (a) $\Rightarrow$ (b): This is already shown in Lemma 2.3.1.
- (b) $\Rightarrow$ (c): Assume $\mathbb{P}(|X| \ge t) \le 2\exp\left(-t^2/c_2^2\right)$. We first observe that we can write

$$\mathbb{E}[|X|^p] = \int_0^\infty pt^{p-1}\mathbb{P}(|X| > t)\,\mathrm{d}t$$

since by working backwards, we have

$$\begin{aligned}
\int_0^\infty pt^{p-1}\mathbb{P}(|X| \ge t)\,\mathrm{d}t &= \int_0^\infty pt^{p-1}\left(\int \mathbb{1}_{t\le|X|}\,\mathrm{d}\mathbb{P}\right)\mathrm{d}t \\
&= \iint_0^\infty pt^{p-1}\mathbb{1}_{t\le|X|}\,\mathrm{d}t\,\mathrm{d}\mathbb{P} \\
&= \iint_0^{|X|} pt^{p-1}\,\mathrm{d}t\,\mathrm{d}\mathbb{P} \\
&= \int |X|^p\,\mathrm{d}\mathbb{P} = \mathbb{E}[|X|^p],
\end{aligned}$$

where the interchanging of the order of integration is given by Tonally's theorem. With this,

we see that

$$\begin{aligned}
\mathbb{E}\left[|X|^p\right] &= \int_0^\infty pt^{p-1}\mathbb{P}(|X| \geq t)\,\mathrm{d}t \\
&\leq \int_0^\infty pt^{p-1} 2\cdot e^{-t^2/c_2^2}\,\mathrm{d}t \\
&= p\cdot c_2^p \int_0^\infty u^{\frac{p}{2}-1}e^{-u}\,\mathrm{d}u && u = t^2/c_2^2 \\
&= p\cdot c_2^p\cdot\Gamma(p/2),
\end{aligned}$$

This implies $(\mathbb{E}\left[|X|^p\right])^{1/p} \leq (pc_2^p\cdot\Gamma(p/2))^{1/p}$. Using the Stirling's formula $\Gamma(x) \leq x^x$, we have

$$(\mathbb{E}\left[|X|^p\right])^{1/p} \leq \left(p\cdot c_2^p\Gamma\left(\frac{p}{2}\right)\right)^{1/p} \leq \left(p\cdot c_2^p\left(\frac{p}{2}\right)^{p/2}\right)^{1/p} \leq \frac{e^{1/e}c_2}{\sqrt{2}}\sqrt{p} =: c_3\sqrt{p}$$

where we use the fact that $\max x^{1/x} = e^{1/e}$.

- (c) ⇒ (d): Assume $(\mathbb{E}\left[|X|^p\right])^{1/p} \leq c_3\sqrt{p}$. Then from the fact that both sides are non-negative and $p \geq 1$, we can safely raise both sides to power of $p$ and get

$$\mathbb{E}\left[|X|^p\right] \leq c_3^p p^{p/2}.$$

From $e^x = 1 + \sum_{p=1}^\infty x^p/p!$, for all $|\lambda| \leq 1/c_4$ for some $c_4 > 0$, we have

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda^2X^2}\right] &= \mathbb{E}\left[1 + \sum_{p=1}^\infty \frac{(\lambda^2X^2)^p}{p!}\right] \\
&= 1 + \sum_{p=1}^\infty \frac{\lambda^{2p}}{p!}\mathbb{E}\left[X^{2p}\right] \\
&\leq 1 + \sum_{p=1}^\infty \frac{(c_3^2\cdot\lambda^2)^p}{p!}\cdot(2p)^p \\
&= 1 + \sum_{p=1}^\infty (2c_3^2\lambda^2)^p\cdot\frac{p^p}{p!} \\
&\leq 1 + \sum_{p=1}^\infty (2c_3^2\lambda^2)^p\frac{p^p}{(p/e)^p} && \text{since } p! \geq (p/e)^p \\
&= \sum_{p=0}^\infty (2ec_3^2\lambda^2)^p \\
&= \frac{1}{1-2ec_3^2\lambda^2}
\end{aligned}$$

if $2e\lambda^2c_3^2 < 1 \Leftrightarrow |\lambda| < \frac{1}{\sqrt{2e}c_3}$. Recall that $1/(1-x) \leq e^{2x}$ for $x \in [0, 1/2]$, we further have

$$\mathbb{E}\left[e^{\lambda^2X^2}\right] \leq e^{4ec_3^2\lambda^2}$$

for all $|\lambda| \leq \min(\frac{1}{\sqrt{2e}c_3}, \frac{1}{2c_3\sqrt{e}}) = \frac{1}{2c_3\sqrt{e}}$,[a] i.e., by letting $c_4 \coloneqq 2c_3\sqrt{e}$,

$$\mathbb{E}\left[e^{\lambda^2X^2}\right] \leq e^{c_4^2\lambda^2}$$

as desired.

- (d) ⇒ (e): Assume that for all $\lambda$ such that $|\lambda| \le 1/c_4$, we have $\mathbb{E}\left[e^{\lambda^2 X^2}\right] \le e^{c_4^2\lambda^2}$. Then, by choosing $c_5 < \infty$ such that $1/c_5^2 \le 1/c_4^2$ (i.e., $\lambda^2 \coloneqq 1/c_5^2$),

$$\mathbb{E}\left[e^{X^2/c_5^2}\right] \le e^{c_4^2/c_5^2} \le 2$$

for large enough $c_5^2$.[b]

- (e) ⇒ (a): Assume that $\mathbb{E}\left[e^{X^2/c_5^2}\right] \le 2$ for some $c_5 < \infty$. Then for all $\lambda \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda X}\right] &= 1 + \mathbb{E}\left[\sum_{p=2}^{\infty} \frac{(\lambda X)^p}{p!}\right] && \text{since } \mathbb{E}[X] = 0 \\
&\le 1 + \frac{\lambda^2}{2}\mathbb{E}\left[X^2 \sum_{p=2}^{\infty} \frac{|\lambda X|^{p-2}}{(p-2)!}\right] && \text{extract } \tfrac{\lambda^2 X^2}{2}\text{, holds since } p \ge 2 \\
&= 1 + \frac{\lambda^2}{2}\mathbb{E}\left[X^2 e^{|\lambda X|}\right] \\
&\le 1 + \frac{\lambda^2}{2}\mathbb{E}\left[X^2 \exp\left(\frac{\lambda^2 c_5^2}{2} + \frac{X^2}{2c_5^2}\right)\right] && ab \le a^2/2 + b^2/2 \\
&= 1 + \frac{\lambda^2}{2} \cdot \exp\left(\frac{\lambda^2 c_5^2}{2}\right) \cdot \mathbb{E}\left[X^2 \exp\left(\frac{X^2}{2c_5^2}\right)\right] \\
&= 1 + \frac{\lambda^2}{2} \cdot \exp\left(\frac{\lambda^2 c_5^2}{2}\right) \cdot 2c_5^2 \cdot \mathbb{E}\left[\frac{X^2}{2c_5^2} \exp\left(\frac{X^2}{2c_5^2}\right)\right] \\
&\le 1 + \lambda^2 c_5^2 \cdot \exp\left(\frac{\lambda^2 c_5^2}{2}\right) \cdot \mathbb{E}\left[e^{X^2/c_5^2}\right] && \text{since } X^2/2c_5^2 \le e^{X^2/2c_5^2} \\
&\le 1 + 2\lambda^2 c_5^2 \cdot \exp\left(\frac{\lambda^2 c_5^2}{2}\right) \\
&\le \left(1 + 2\lambda^2 c_5^2\right) \exp\left(\frac{\lambda^2 c_5^2}{2}\right) \\
&\le \exp\left(2\lambda^2 c_5^2 + \frac{\lambda^2 c_5^2}{2}\right) && 1 + x \le e^x \\
&= \exp\left(\frac{5c_5^2}{2}\lambda^2\right).
\end{aligned}$$

By letting $c_1^2 \coloneqq 5c_5^2/2$, we recover $\mathbb{E}\left[e^{\lambda X}\right] \le e^{c_1^2\lambda^2}$ for all $\lambda \in \mathbb{R}$.

■

[a] We omit the fact that we require a strict $<$ in the first case.
[b] Which corresponds to small enough $|\lambda| = 1/c_5$ hence the conditions are consistent.

**Note.** It's possible to assume $c_1, \dots, c_5 = 1$, which significantly simplify the proof.

Before proving Lemma 2.3.2 with the correct constant, we need a small lemma.

**Lemma A.1.2.** For any bounded random variable $Z \in [a, b]$,

$$\operatorname{Var}[Z] \le \frac{(b-a)^2}{4}.$$

**Proof.** Since

$$\operatorname{Var}[Z] = \operatorname{Var}\left[Z - \frac{a+b}{2}\right] \le \mathbb{E}\left[\left(Z - \frac{a+b}{2}\right)^2\right] \le \frac{(b-a)^2}{4}.$$

■

We can then see the correct proof.

**Lemma A.1.3** (Lemma 2.3.2). Given $X \in [a,b]$ such that $\mathbb{E}[X] = 0$. Then

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

for all $\lambda \in \mathbb{R}$, i.e., $X \in \mathrm{Subg}((b-a)^2/4)$.

**Proof.** We first define $\psi(\lambda) = \ln \mathbb{E}\left[e^{\lambda X}\right]$, and compute

$$\psi'(\lambda) = \frac{\mathbb{E}\left[Xe^{\lambda X}\right]}{\mathbb{E}\left[e^{\lambda X}\right]}, \quad \psi''(\lambda) = \frac{\mathbb{E}\left[X^2 e^{\lambda X}\right]}{\mathbb{E}\left[e^{\lambda X}\right]} - \left(\frac{\mathbb{E}\left[Xe^{\lambda X}\right]}{\mathbb{E}\left[e^{\lambda X}\right]}\right)^2.$$

Now, observe that $\psi''$ is the variance under the law of $X$ re-weighted by $\frac{e^{\lambda X}}{\mathbb{E}[e^{\lambda X}]}$, i.e., by a change of measure, consider a new distribution $\mathbb{P}_\lambda$ (w.r.t. the original distribution $\mathbb{P}$ of $X$) as

$$\mathrm{d}\mathbb{P}_\lambda(x) \coloneqq \frac{e^{\lambda X}}{\mathbb{E}_{\mathbb{P}}\left[e^{\lambda X}\right]} \mathrm{d}\mathbb{P}(x),$$

then

$$\psi'(\lambda) = \frac{\mathbb{E}_{\mathbb{P}}\left[Xe^{\lambda X}\right]}{\mathbb{E}_{\mathbb{P}}\left[e^{\lambda X}\right]} = \int \frac{xe^{\lambda x}}{\mathbb{E}_{\mathbb{P}}\left[e^{\lambda X}\right]} \mathrm{d}\mathbb{P}(x) = \mathbb{E}_{\mathbb{P}_\lambda}[X]$$

and

$$\psi''(\lambda) = \frac{\mathbb{E}_{\mathbb{P}}\left[X^2 e^{\lambda X}\right]}{\mathbb{E}_{\mathbb{P}}\left[e^{\lambda X}\right]} - \left(\frac{\mathbb{E}_{\mathbb{P}}\left[Xe^{\lambda X}\right]}{\mathbb{E}_{\mathbb{P}}\left[e^{\lambda X}\right]}\right)^2 = \mathbb{E}_{\mathbb{P}_\lambda}\left[X^2\right] - \mathbb{E}_{\mathbb{P}_\lambda}[X]^2 = \mathrm{Var}_{\mathbb{P}_\lambda}[X].$$

From Lemma A.1.2, since $X$ under the new distribution $\mathbb{P}_\lambda$ is still bounded between $a$ and $b$,

$$\psi''(\lambda) = \mathrm{Var}_{\mathbb{P}_\lambda}[X] \le \frac{(b-a)^2}{4}.$$

Then by Taylor's theorem, there exists some $\widetilde{\lambda} \in [0,\lambda]$ such that

$$\psi(\lambda) = \psi(0) + \psi'(0)\lambda + \frac{1}{2}\psi''(\widetilde{\lambda})\lambda^2 = \frac{1}{2}\psi''(\widetilde{\lambda})\lambda^2$$

since $\psi(0) = \psi'(0) = 0$. By bounding $\psi''(\widetilde{\lambda})\lambda^2/2$, we finally have

$$\ln \mathbb{E}\left[e^{\lambda X}\right] = \psi(\lambda) \le \frac{1}{2} \cdot \frac{(b-a)^2}{4}\lambda^2 = \lambda^2 \frac{(b-a)^2}{8},$$

or equivalently,

$$\mathbb{E}\left[e^{\lambda X}\right] \le \exp\left(\lambda^2 \frac{(b-a)^2}{8}\right).$$

■

**Remark.** Comparing the proof given in Lemma 2.3.2, we see that the correct proof is actually easier.

We now prove Lemma 2.3.4.

**Lemma A.1.4.** Let $X_1, \dots, X_n \sim \mathrm{Subg}(\sigma_i^2)$, not necessary independent. Then for some absolute constant $c > 0$,

$$\mathbb{E}\left[\max_i |X_i|\right] \le c\sqrt{\log n} \max_{1 \le i \le n} \sigma_i.$$

**Proof.** We see that for any $\lambda > 0$,

$$\begin{aligned}
\mathbb{E}\left[\max_{i\in[n]} X_i\right] &= \frac{1}{\lambda}\mathbb{E}\left[\ln\exp\left(\lambda \max_{i\in[n]} X_i\right)\right] \\
&\le \frac{1}{\lambda}\ln\mathbb{E}\left[\exp\left(\lambda \max_{i\in[n]} X_i\right)\right] && \text{Jensen's inequality} \\
&= \frac{1}{\lambda}\ln\mathbb{E}\left[\max_{i\in[n]} e^{sX_i}\right] \\
&\le \frac{1}{\lambda}\ln\mathbb{E}\left[\sum_{i=1}^{n} e^{\lambda X_i}\right] \\
&\le \frac{1}{\lambda}\ln\sum_{i=1}^{n}\exp\left(\frac{\sigma_i^2\lambda^2}{2}\right) \\
&\le \frac{1}{\lambda}\ln\left(n\cdot\exp\left(\frac{\lambda^2\cdot\max_i\sigma_i^2}{2}\right)\right) \\
&= \frac{\ln n}{\lambda} + \frac{\lambda\max_i\sigma_i^2}{2}.
\end{aligned}$$

Now, take $\lambda = \sqrt{2\ln n/\max_i\sigma_i^2}$, we have

$$\mathbb{E}\left[\max_{i\in[n]} X_i\right] \le \sqrt{\frac{\ln n\cdot\max_i\sigma_i^2}{2}} + \frac{\sqrt{2\ln n\cdot\max_i\sigma_i^2}}{2} = \sqrt{2}\cdot\sqrt{\ln n}\cdot\max_i\sigma_i.$$

Finally, apply this result to $\{Z_i\}_{i\in[2n]}$ such that $Z_i = X_i$ and $Z_{i+n} = -X_i$ for $i\in[n]$, then

$$\mathbb{E}\left[\max_{i\in[n]}|X_i|\right] = \mathbb{E}\left[\max_{i\in[2n]} Z_i\right] \le \sqrt{2}\cdot\sqrt{\ln 2n}\cdot\max_i\sigma_i,$$

which is the desired result by setting $c \coloneqq \sqrt{2\ln 2n}/\sqrt{\ln n}$. ■

### A.1.2 Bounded Difference Concentration Inequality

As mentioned, in the lecture, we did not prove McDiarmid's inequality in fact. To properly prove it, we will need the tool of martingale decomposition and Azuma-Hoeffding inequality. We start by proving the latter. Consider first the following definition.

**Definition A.1.1** (Martingale difference sequence)**.** A *martingale difference sequence* is a sequence of random variables $\Delta_1,\dots$ such that $\mathbb{E}\left[\Delta_i \mid \Delta_{i-1}\right] = 0$ for all $i$.

**Theorem A.1.1** (Azuma-Hoeffding inequality)**.** Suppose $\mathcal{F}_0,\mathcal{F}_1,\dots,\mathcal{F}_n$ are increasing $\sigma$-fields, and let $X_i\in\mathcal{F}_i$ measurable and $\mathbb{E}\left[X_i \mid \mathcal{F}_{i-1}\right] = 0$ almost surely for all $i = 1,\dots,n$, i.e., $X_1, X_2,\dots,X_n$ is a martingale difference sequence. Suppose also that $|X_i| \le c_i$ almost surely for all $i = 1,\dots,n$. Then, the Hoeffding's inequality holds for this martingale difference sequence, i.e.,

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \ge t\right) \le \exp\left(-\frac{t^2}{2\sum_i c_i^2}\right),$$

and the same bound holds for the left tail.

**Proof.** First, by MGF trick, we have that for any $\lambda > 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \ge t\right) \le \frac{\mathbb{E}\left[\exp(\lambda\sum_i X_i)\right]}{\exp(\lambda t)}.$$

Note that since $|X_i| \le c_i$, we have $X_i \in [-c_i, c_i]$. Now,

$$\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} X_i\right)\right] = \mathbb{E}\left[\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} X_i\right)\middle| \mathcal{F}_{n-1}\right]\right]$$

since $\exp\left(\lambda \sum_{i=1}^{n-1} X_i\right)$ is $\mathcal{F}_{n-1}$-measurable,

$$= \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n-1} X_i\right)\mathbb{E}\left[\exp(\lambda X_n) \mid \mathcal{F}_{n-1}\right]\right]$$

from Lemma 2.3.2 and $|X_i| \le c_i$, $\mathbb{E}\left[\exp(\lambda X_n) \mid \mathcal{F}_{n-1}\right] \le \exp\left(\lambda^2 (2c_n)^2/8\right) = \exp\left(\lambda^2 c_n^2/2\right)$,

$$\le \exp\left(\frac{\lambda^2 c_n^2}{2}\right)\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n-1} X_i\right)\right].$$

We see that this can be repeatedly apply to the last $X_i$ and get

$$\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} X_i\right)\right] \le \exp\left(\frac{\lambda^2}{2}\sum_{i=1}^{n} c_i^2\right),$$

i.e., $\sum_i X_i \in \mathrm{Subg}(\sum_i c_i^2)$. From (one-sided) Hoeffding's inequality,

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \ge t\right) \le \exp\left(\frac{-t^2}{2\sum_i c_i^2}\right).$$

■

The above is actually a special symmetric version since we have a two-sided bound for $X_i$, i.e., we assume $|X_i| \le c_i$. However, if the known bound is asymmetric, e.g., $a_i \le X_i \le b_i$, to use Azuma-Hoeffding inequality, one would need to choose $c_i = \max(|a_i|, |b_i|)$, which might not be optimal. A slight change can be made as follows.

**Corollary A.1.1** (General Azuma-Hoeffding inequality). Suppose $\mathcal{F}_0, \mathcal{F}_1, \ldots, \mathcal{F}_n$ are increasing $\sigma$-fields, and let $X_i \in \mathcal{F}_i$ measurable and $\mathbb{E}\left[X_i \mid \mathcal{F}_{i-1}\right] = 0$ almost surely for all $i = 1, \ldots, n$, i.e., $X_1, X_2, \ldots, X_n$ is a martingale difference sequence. Suppose also that $A_i \le X_i \le B_i$ almost surely for all $i = 1, \ldots, n$ such that $B_i - A_i \le c_i$ almost surely. Then, the Hoeffding's inequality holds for this martingale difference sequence, i.e.,

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \ge t\right) \le \exp\left(-\frac{2t^2}{\sum_i c_i^2}\right),$$

and the same bound holds for the left tail.

Then, we introduce the tool of martingale decomposition.

**Definition A.1.2** (Martingale decomposition). Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$ on $\chi$. Let $f : \chi^n \to \mathbb{R}$ satisfying the bounded-difference property with parameters $c_1, \ldots, c_n$. Then the *martingale decomposition* of $f$ is defined as

$$\begin{aligned} f(X_1, \ldots, X_n) - \mathbb{E}\left[f(X_1, \ldots, X_n)\right] &= Y_n - Y_0 \\ &= (Y_n - Y_{n-1}) + (Y_{n-1} - Y_{n-2}) + \cdots + (Y_1 - Y_0) \\ &=: \Delta_n + \Delta_{n-1} + \cdots + \Delta_1, \end{aligned}$$

where $Y_i = \mathbb{E}\left[f(X_1, \ldots, X_n) \mid X_1, \ldots, X_i\right]$ and $\Delta_i \coloneqq Y_i - Y_{i-1}$.

**Note.** For any integrable random variable $Y$, $Y_i = \mathbb{E}\left[Y \mid X_1, \ldots, X_i\right]$ form a martingale difference sequence if centered, i.e., $\Delta_i = Y_i - Y_{i-1}$.

Then, we can show the McDiarmid's inequality.

**Theorem A.1.2** (McDiarmid's inequality (Theorem 2.5.1)). Let $X_1, \dots, X_n$ be i.i.d. random variables on $\chi$, and let $f\colon \chi^n \to \mathbb{R}$ satisfying the bounded difference property with parameters $c_1, \dots, c_n$. Then for any $t > 0$,

$$\mathbb{P}(f(X_1, \dots, X_n) - \mathbb{E}\left[f(X_1, \dots, X_n)\right] \geq t) \leq \exp\left(\frac{-2t^2}{\sum_i c_i^2}\right).$$

The same bound holds for the left tail.

**Proof.** Consider the martingale decomposition of $f$ such that $f(X_1, \dots, X_n) - \mathbb{E}\left[f(X_1, \dots, X_n)\right] = \sum_{i=1}^n \Delta_i$. Denote $X = (X_1, \dots, X_n)$, and recall that

$$h_i(x) = \mathbb{E}\left[f(X) \mid X_1, \dots, X_{i-1}\right] = \mathbb{E}\left[f(x_1, \dots, x_{i-1}, x, X_{i+1}, \dots, X_n)\right],$$

and

$$Y_i = \mathbb{E}\left[f(X) \mid X_1, \dots, X_i\right] = \mathbb{E}\left[f(x_1, \dots, x_i, X_{i+1}, \dots, X_n)\right].$$

Now, define random variables

$$A_i \coloneqq \inf_{x \in \chi} h_i(x) - Y_{i-1}, \quad B_i \coloneqq \sup_{x \in \chi} h_i(x) - Y_{i-1}.$$

It's obvious that $B_i \geq A_i$ almost everywhere for all $i = 1, \dots, n$. Moreover, we have

$$\Delta_i - A_i = (Y_i - Y_{i-1}) - (\inf_{x \in \chi} h_i(x) - Y_{i-1}) = Y_i - \inf_{x \in \chi} h_i(x) \geq 0$$

almost everywhere, and similarly $\Delta_i - B_i \leq 0$, hence $A_i \leq \Delta_i \leq B_i$. Observe that

$$\begin{aligned}
B_i - A_i &= \sup_{x \in \chi} h_i(x) - \inf_{y \in \chi} h_i(y) \\
&= \sup_{x, y \in \chi} h_i(x) - h_i(y) \\
&= \sup_{x, y \in \chi} \mathbb{E}\left[f(x_1, \dots, x_{i-1}, x, X_{i+1}, \dots, X_n) - f(x_1, \dots, x_{i-1}, y, X_{i+1}, \dots, X_n)\right] \\
&\leq \mathbb{E}\left[\sup_{x, y \in \chi} \left|f(x_1, \dots, x_{i-1}, x, X_{i+1}, \dots, X_n) - f(x_1, \dots, x_{i-1}, y, X_{i+1}, \dots, X_n)\right|\right] \\
&\leq c_i.
\end{aligned}$$

Then, by applying the general Azuma-Hoeffding inequality to $\sum_{i=1}^n \Delta_i$, we have

$$\mathbb{P}(f(X) - \mathbb{E}\left[f(X)\right] \geq t) = \mathbb{P}\left(\sum_{i=1}^n \Delta_i \geq t\right) \leq \exp\left(\frac{-2t^2}{\sum_i c_i^2}\right).$$

■

Finally, we provide a proof for the Efron-Stein inequality.

**Theorem A.1.3** (Efron-Stein inequality (Theorem 2.5.2)). Let $X_1, \dots, X_n$ be independent random variables, and $X_1', \dots, X_n'$ be i.i.d. copies of $X_i$'s. Then

$$\operatorname{Var}\left[f(X)\right] \leq \frac{1}{2} \sum_{i=1}^n \mathbb{E}\left[\left(f(X) - f(X^{[i]})\right)^2\right].$$

**Proof.** Denote $X = (X_1, \dots, X_n)$, $X' = (X_1', \dots, X_n')$, $X_{[i:j]} = (X_i, \dots, X_j)$, and

$$\mathbb{E}_i\left[f(X)\right] = \mathbb{E}\left[f(X) \mid X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n\right] = \mathbb{E}\left[f(X) \mid X_{[1:i-1]}, X_{[i+1:n]}\right].$$

Define $\Delta_i$ as $\Delta_i = Y_i - Y_{i-1}$ where $Y_i$ is defined the same as in martingale decomposition. Hence,

$f(X) - \mathbb{E}[f(X)] = \sum_{i=1}^n \Delta_i$. Now, observe that

$$\mathrm{Var}[f(X)] = \mathbb{E}\left[(f(X) - \mathbb{E}[f(X)])^2\right] = \mathbb{E}\left[\left(\sum_{i=1}^n \Delta_i\right)^2\right] = \mathbb{E}\left[\sum_{i=1}^n \Delta_i^2\right] + 2\mathbb{E}\left[\sum_{i>j} \Delta_i \Delta_j\right].$$

Since $\mathbb{E}[XY] = \mathbb{E}[\mathbb{E}[XY \mid Y]] = \mathbb{E}[Y\mathbb{E}[X \mid Y]]$, $\mathbb{E}[\Delta_j \Delta_i] = \mathbb{E}\left[\Delta_i \mathbb{E}\left[\Delta_j \mid X_{[1:i]}\right]\right]$. But since for $i > j$, $\mathbb{E}\left[\Delta_j \mid X_{[1:i]}\right] = 0$, we have

$$\mathrm{Var}[f(X)] = \mathbb{E}\left[\sum_{i=1}^n \Delta_i^2\right] = \sum_{i=1}^n \mathbb{E}\left[\Delta_i^2\right].$$

Now, expand everything,

$$\begin{aligned}
\mathrm{Var}[f(X)] &= \sum_{i=1}^n \mathbb{E}\left[\Delta_i^2\right] \\
&= \sum_{i=1}^n \mathbb{E}_{X_{[1:i]}}\left[\left(\mathbb{E}_{X_{[i+1:n]}}\left[f(X) \mid X_{[1:i]}\right] - \mathbb{E}_{X_{[i:n]}}\left[f(X) \mid X_{[1:i-1]}\right]\right)^2\right] \\
&= \sum_{i=1}^n \mathbb{E}_{X_{[1:i]}}\left[\left(\mathbb{E}_{X_{[i+1:n]}}\left[f(X) \mid X_{[1:i]}\right] - \mathbb{E}_{X_{[i+1:n]}}\left[\mathbb{E}_{X_i}\left[f(X) \mid X_{[1:i-1]}\right]\right]\right)^2\right] \\
&\le \sum_{i=1}^n \mathbb{E}_{X_{[1:i]}, X_{[i+1:n]}}\left[\left(f(X) - \mathbb{E}_{X_i}[f(X)]\right)^2\right]
\end{aligned}$$

where the last line follows from Jensen's inequality with the fact that $x^2$ is convex. Using our notation, we write

$$\mathrm{Var}[f(X)] \le \sum_{i=1}^n \mathbb{E}\left[\left(f(X) - \mathbb{E}_i[f(X)]\right)^2\right].$$

Finally, note that for two i.i.d. samples $x, y \overset{\text{i.i.d.}}{\sim} \mathbb{P}$,

$$\mathbb{E}\left[(x-y)^2\right] = \mathbb{E}\left[x^2 + y^2 - 2xy\right] = 2\mathbb{E}\left[x^2\right] - 2(\mathbb{E}[x])^2 \Rightarrow \mathrm{Var}[x] = \mathbb{E}\left[\frac{1}{2}(x-y)^2\right].$$

This implies that

$$\mathbb{E}_i\left[\left(f(X) - \mathbb{E}_i[f(X)]\right)^2\right] = \frac{1}{2}\mathbb{E}_i\left[(f(X) - f(X_1, \ldots, X_i', \ldots, X_n))^2\right],$$

thus we have the final inequality

$$\mathrm{Var}[f(X)] \le \frac{1}{2}\sum_{i=1}^n \mathbb{E}\left[\left(f(X) - f(X_1, \ldots, X_i', \ldots, X_n)\right)^2\right].$$

■

**Remark.** Equivalently, one can write the bound of Efron-Stein inequality as[a]

$$\mathrm{Var}[f(X)] \le \sum_{i=1}^n \mathbb{E}\left[\left(f(X) - f(X_1, \ldots, X_i', \ldots, X_n)\right)_+^2\right].$$

[a] Where $(x)_+$ means $\max(0, x)$, so we're avoiding double-counting.

We have an additional corollary for Efron-Stein inequality.

**Corollary A.1.2.** Let $X_1, \ldots, X_n$ be independent random variables. If the function $f$ satisfies the

bounded-difference property with parameters $c_1, \dots, c_n$, then

$$\operatorname{Var}\left[f(X_1, \dots, X_n)\right] \le \frac{1}{4}\sum_{i=1}^{n} c_i^2.$$

**Proof.** From Lemma A.1.2,

$$\mathbb{E}\left[\left(f(X) - \mathbb{E}_i\left[f(X)\right]\right)^2\right] = \operatorname{Var}_i\left[f(X)\right] \le \frac{c_i^2}{4},$$

hence $\operatorname{Var}\left[f(X)\right] \le \frac{1}{4}\sum_{i=1}^{n} c_i^2$. ■

## A.2 Expected Supremum of Empirical Process

### A.2.1 Metric Entropy Method

We first show the proof of Theorem 3.3.1. Recall the following.

**As previously seen.** Let $d(f,g) = \sup_{x\in[0,1]} |f(x) - g(x)|$, then $(\mathcal{S}_\alpha, d)$ is a pseudo-metric space.

The proof is rather long, and we first state and prove some general facts about functions in $\mathcal{S}_\alpha$. First, recall the Taylor's theorem.

**Theorem A.2.1** (Taylor's theorem). Let $k \ge 1$ be an integer and let $f: \mathbb{R} \to \mathbb{R}$ be $k$ times differentiable at the point $a \in \mathbb{R}$. Then there exists some real number $\xi \in (a, x)$ such that

$$f(x) = \sum_{i=0}^{k-1} \frac{(x-a)^i}{i!} f^{(i)}(x) + \frac{f^{(k)}(\xi)}{k!}(x-a)^k.$$

**Lemma A.2.1.** For every $f \in \mathcal{S}_\alpha$ and $x, x+h \in (0,1)$,

$$f(x+h) = \sum_{k=0}^{\beta} \frac{h^k}{k!} f^{(k)}(x) + R_f(x,h)$$

where the remainder term $R_f(x,h)$ satisfies $|R_f(x,h)| \le |h|^\alpha/\beta!$.

**Proof.** By directly applying Taylor's theorem in our case, we see that

$$f(x+h) = \sum_{k=0}^{\beta-1} \frac{h^k}{k!} f^{(k)}(x) + \frac{f^{(\beta)}(\xi)}{\beta!} h^\beta$$

for some $\xi$ between $x$ and $x+h$. We can then rewrite this as

$$f(x+h) = \sum_{k=0}^{\beta} \frac{h^k}{k!} f^{(k)}(x) + \underbrace{\frac{h^\beta (f^{(\beta)}(\xi) - f^{(\beta)}(x+h))}{\beta!}}_{:= R_f(x,h)}.$$

To show that $|R_f(x,h)| \le |h|^\alpha/\beta!$, we see that

$$|R_f(x,h)| = \frac{|h|^\beta}{\beta!}|f^{(\beta)}(\xi) - f^{(\beta)}(x+h)| \le \frac{|h|^\beta}{\beta!} \cdot |x+h-\xi|^{\alpha-\beta} \le \frac{|h|^\beta}{\beta!} \cdot |h|^{\alpha-\beta} = \frac{|h|^\alpha}{\beta!}.$$

■

**Lemma A.2.2.** For every $f \in \mathcal{S}_\alpha$ and $0 \le i \le \beta$, the derivative $g = f^{(i)}$ belongs to $\mathcal{S}_{\alpha-i}$ and hence

$$f^{(i)}(x+h) = \sum_{k=0}^{\beta-i} \frac{h^k}{k!} f^{(i+k)}(x) + R_{f^{(i)}}(x,h)$$

with $|R_{f^{(i)}}(x,h)| \le |h|^{\alpha-i}/(\beta-i)!$ whenever $x, x+h \in (0,1)$.

**Proof.** For $f \in \mathcal{S}_\alpha$ and $0 \le i \le \beta$, we know that $g = f^{(i)}$ satisfies

- $g$ is $(\beta-i)$-times differentiable, and hence it is continuous on $[0,1]$ (if $i < \beta$);[a]
- $|g^{(k)}| \le 1$ for all $k = 0, \dots, \beta - i$;
- $|g^{(\beta-i)}(x) - g^{(\beta-i)}(y)| \le |x-y|^{(\alpha-i)-(\beta-i)}$ for all $x, y \in [0,1]$.

Hence, $g \in \mathcal{S}_{\alpha-i}$. Then from Lemma A.2.1, we have

$$f^{(i)}(x+h) = g(x+h) = \sum_{k=0}^{\beta-i} \frac{h^k}{k!} g^{(k)}(x) + R_g(x,h) = \sum_{k=0}^{\beta-i} \frac{h^k}{k!} f^{(i+k)}(x) + R_{f^{(i)}}(x,h)$$

where $|R_{f^{(i)}}(x,h)| = |R_g(x,h)| \le |h|^{\alpha-i}/(\beta-i)!$. ■

[a] When $i = \beta$, $g \in \mathcal{S}_0$ obviously.

**Lemma A.2.3.** Fix $\epsilon > 0$ and $x \in (0,1)$. Suppose that two functions $f$ and $g$ in $\mathcal{S}_\alpha$ satisfy

$$|f^{(k)}(x) - g^{(k)}(x)| \le \epsilon^{1-k/\alpha}$$

for all $k = 0, 1, \dots, \beta$. Then for every $h \in \mathbb{R}$ such that $|h| \le \epsilon^{1/\alpha}$ and $x + h \in (0,1)$,

$$|f(x+h) - g(x+h)| \le C(\alpha)\epsilon$$

where $C(\alpha) = \sum_{k=0}^{\beta} 1/k! + 2/\beta!$.

**Proof.** Given a fixed $\epsilon > 0$, we know that for every $h \in \mathbb{R}$ with $|h| \le \epsilon^{1/\alpha}$ and $x, x+h \in (0,1)$,

$$\begin{aligned}
|f(x+h) - g(x+h)| &= \left| \left( \sum_{k=0}^{\beta} \frac{h^k}{k!} f^{(k)}(x) + R_f(x,h) \right) - \left( \sum_{k=0}^{\beta} \frac{h^k}{k!} g^{(k)}(x) + R_g(x,h) \right) \right| \\
&\le \left| \sum_{k=0}^{\beta} \frac{h^k}{k!} (f^{(k)}(x) - g^{(k)}(x)) \right| + |R_f(x,h) - R_g(x,h)| \\
&\le \left| \sum_{k=0}^{\beta} \frac{h^k}{k!} \epsilon^{1-(k/\alpha)} \right| + |R_f(x,h) - R_g(x,h)| \\
&\le \sum_{k=0}^{\beta} \frac{|h|^k}{k!} \epsilon^{1-(k/\alpha)} + \left| R_f(x,h) - \frac{|h|^\alpha}{\beta!} \right| + \left| R_g(x,h) - \frac{|h|^\alpha}{\beta!} \right| \\
&\le \sum_{k=0}^{\beta} \frac{|h|^k}{k!} \epsilon^{1-(k/\alpha)} + 2\frac{|h|^\alpha}{\beta!} \\
&\le \sum_{k=0}^{\beta} \frac{\epsilon^{k/\alpha}}{k!} \epsilon^{1-(k/\alpha)} + 2\frac{\epsilon}{\beta!} \\
&= \sum_{k=0}^{\beta} \frac{1}{k!} \epsilon + 2\frac{\epsilon}{\beta!} =: C(\alpha)\epsilon
\end{aligned}$$

for $C(\alpha) = \sum_{k=0}^{\beta} 1/k! + 2/\beta!$. ■

**Theorem A.2.2** (Theorem 3.3.1). There exists constants $c_1, c_2 > 0$ only depend on $\alpha$ such that for all $\epsilon > 0$,

$$\exp\left(c_2\epsilon^{-1/\alpha}\right) \le M(\mathcal{S}_\alpha, d, \epsilon) \le \exp\left(c_1\epsilon^{-1/\alpha}\right).$$

**Proof.** We start by showing the upper-bound. In the rest of the proof, let $C(\alpha)$ be a generic constant depending only on $\alpha$, so it might be distinct from $C(\alpha)$ defined above.

**Claim.** Let $x_1 < \cdots < x_s$ be a maximal $\epsilon^{1/\alpha}$ separated set in $(0,1)$.[a] For each $f_0 \in \mathcal{S}_\alpha$, consider the following subset of $\mathcal{S}_\alpha$

$$\mathcal{G}(f_0) \coloneqq \left\{ f \in \mathcal{S}_\alpha : \left\lfloor \frac{f^{(k)}(x_i)}{\epsilon^{1-k/\alpha}} \right\rfloor = \left\lfloor \frac{f_0^{(k)}(x_i)}{\epsilon^{1-k/\alpha}} \right\rfloor \text{ for all } 1 \le i \le s \text{ and } 0 \le k \le \beta \right\}.$$

Then the number of distinct sets $\mathcal{G}(f_0)$ as $f_0$ ranges over $\mathcal{S}_\alpha$ is an upper-bound on $N(\mathcal{S}_\alpha, d, C(\alpha)\epsilon)$, and hence $N(\mathcal{S}_\alpha, d, C(\alpha)\epsilon)$ is bounded from above by the cardinality of the set $I$ defined as

$$I \coloneqq \left\{ \left( \left\lfloor \frac{f^{(k)}(x_i)}{\epsilon^{1-k/\alpha}} \right\rfloor, i = 1, \dots, s \text{ and } k = 0, \dots, \beta \right) : f \in \mathcal{S}_\alpha \right\}.$$

[a] In the usual Euclidean metric on $\mathbb{R}$.

**Proof.** Let $N$ be a minimal $C(\alpha)\epsilon$-net of $(\mathcal{S}_\alpha, d)$ such that $|N| = N(\mathcal{S}_\alpha, d, C(\alpha)\epsilon)$. We want to show that $|\{\mathcal{G}(f)\}_{f\in\mathcal{S}_\alpha}| \ge N(\mathcal{S}_\alpha, d, C(\alpha)\epsilon)$. It suffices to show that the set of representatives $R$, one from each $\mathcal{G}(f)$, forms a $C(\alpha)\epsilon$-net. If this is the case, then $|R| \ge |N|$ by the minimality of $N(\mathcal{S}_\alpha, d, C(\alpha)\epsilon)$. For two representative $f \ne g \in R$ picked from $\mathcal{G}(f) \ne \mathcal{G}(g)$, respectively, there exists $1 \le i \le s$ and $0 \le k \le \beta$ such that $\left\lfloor \frac{f^{(k)}(x_i)}{\epsilon^{1-(k/\alpha)}} \right\rfloor \ne \left\lfloor \frac{g^{(k)}(x_i)}{\epsilon^{1-(k/\alpha)}} \right\rfloor$. Choosing the closest $g$ w.r.t. $f$ in terms of the difference, i.e., $\left| \left\lfloor \frac{f^{(k)}(x_i)}{\epsilon^{1-(k/\alpha)}} \right\rfloor - \left\lfloor \frac{g^{(k)}(x_i)}{\epsilon^{1-(k/\alpha)}} \right\rfloor \right| \ge 1$. Observe that the difference is actually attainable by 1 as $g$ can range over the entire $\mathcal{S}_\alpha$, hence, there exists a $g$ such that $|f^{(k)}(x_i) - g^{(k)}(x_i)| \le \epsilon^{1-(k/\alpha)}$. Then from Lemma A.2.3, for $x \in [x_i - \epsilon^{1/\alpha}, x_i + \epsilon^{1/\alpha}]$, $|f(x) - g(x)| \le C(\alpha)\epsilon$. Since this holds for every $x_i$ (which is separated by $\epsilon^{1/\alpha}$ exactly), this bounds extends for all $x \in (0,1)$, i.e., $d(f,g) = \sup_{0<x<1} |f(x) - g(x)| \le C(\alpha)\epsilon$, hence $R$ is a $C(\alpha)\epsilon$-net, implying $|\{\mathcal{G}(f)\}_{f\in\mathcal{S}_\alpha}| \ge N(\mathcal{S}_\alpha, d, C(\alpha)\epsilon)$.

It's now easy to see that $N(\mathcal{S}_\alpha, d, C(\alpha)\epsilon) \le |I|$ as for every $\mathcal{G}(f)$, there exists a one-to-one corresponding element in $I$ defined exactly by $\left( \left\lfloor \frac{f^{(k)}(x_i)}{\epsilon^{1-(k/\alpha)}} \right\rfloor \right)_{\substack{i=1,\dots,s \\ k=0,\dots,\beta}}$, hence $|I| = |\{\mathcal{G}(f)\}_{f\in\mathcal{S}_\alpha}| \ge N(\mathcal{S}_\alpha, d, C(\alpha)\epsilon)$. ⊛

**Claim.** The number of possible vectors $(\lfloor f^{(k)}(x_1)/\epsilon^{1-k/\alpha} \rfloor, k = 0, \dots, \beta)$ as $f$ ranges over $\mathcal{S}_\alpha$ is bounded from above by $C(\alpha)\epsilon^{-\beta-1}$.

**Proof.** We first note that as $f \in \mathcal{S}_\alpha$, $|f^{(k)}(x_1)| \le 1$ for all $k = 0, \dots, \beta$, hence for a fixed $k$, the number of possible vale of $\left\lfloor \frac{f^{(k)}(x_1)}{\epsilon^{1-(k/\alpha)}} \right\rfloor$ among different $f \in \mathcal{S}_\alpha$ is bounded above by $2 \cdot \epsilon^{-(1-(k/\alpha))} = 2\epsilon^{k/\alpha - 1}$. Hence, the number of possible vectors is upper-bounded by

$$\prod_{k=0}^{\beta} 2\epsilon^{k/\alpha-1} = 2^{\beta+1} \cdot \epsilon^{\sum_{k=0}^{\beta}(k/\alpha-1)} = 2^{\beta+1}\epsilon^{\frac{\beta^2+\beta}{2\alpha}} \cdot \epsilon^{-(\beta+1)} \le 2^{\beta+1}\epsilon^{-\beta-1}$$

as $\epsilon^{\frac{\beta^2+\beta}{2\alpha}} \le 1$.[a] By setting $2^{\beta+1} = C(\alpha)$, we have $\le C(\alpha)\epsilon^{-\beta-1}$. ⊛

[a] This is easy to see when $\alpha > 1$, $\frac{\beta^2+\beta}{2\alpha} > 0$, $\ln \epsilon^{\frac{\beta^2+\beta}{2\alpha}} \le 0 \Rightarrow \ln \epsilon \le 0$, which is true sine $\epsilon \in (0,1)$. When $\alpha \in (0,1)$, $\beta = 0$, the inequality still holds.

**Claim.** Fix $f \in \mathcal{S}_\alpha$, and let $A_k \coloneqq \left\lfloor \frac{f^{(k)}(x_1)}{\epsilon^{1-(k/\alpha)}} \right\rfloor$ for $k = 0, \dots, \beta$. Then for all $i = 0, \dots, \beta$,

$$\left| f^{(i)}(x_2) - \sum_{k=0}^{\beta-i} \frac{(x_2 - x_1)^k}{k!} A_{i+k} \epsilon^{1-(i+k)/\alpha} \right| \le C(\alpha)\epsilon^{1-i/\alpha}.$$

**Proof.** Fix $f \in \mathcal{S}_\alpha$, and let $A_k \coloneqq \left\lfloor \frac{f^{(k)}(x_1)}{\epsilon^{1-(k/\alpha)}} \right\rfloor$ for $k = 0, \dots, \beta$. From Lemma A.2.2, for all $i = 0, \dots, \beta$,

$$f^{(i)}(x_2) = \sum_{k=0}^{\beta-i} \frac{(x_2 - x_1)^k}{k!} f^{(i+k)}(x_1) + R_{f^{(i)}}(x_1, x_2 - x_1)$$

with $|R_{f^{(i)}}(x_1, x_2 - x_1)| \le |x_2 - x_1|^{\alpha-i}/(\beta - i)!$. Rearranging, with

$$f^{(i)}(x_2) - \sum_{k=0}^{\beta-i} \frac{(x_2 - x_1)^k}{k!} f^{(i+k)}(x_1) = R_{f^{(i)}}(x_1, x_2 - x_1),$$

we then have [a]

$$\begin{aligned}
&\left| f^{(i)}(x_2) - \sum_{k=0}^{\beta-i} \frac{(x_2 - x_1)^k}{k!} \epsilon^{1-(i+k)/\alpha} \cdot A_{i+k} \right| \\
&= \left| f^{(i)}(x_2) - \sum_{k=0}^{\beta-i} \frac{(x_2 - x_1)^k}{k!} \epsilon^{1-(i+k)/\alpha} \cdot \left\lfloor \frac{f^{(i+k)}(x_1)}{\epsilon^{1-(i+k)/\alpha}} \right\rfloor \right| \\
&\le \left| f^{(i)}(x_2) - \sum_{k=0}^{\beta-i} \frac{(x_2 - x_1)^k}{k!} f^{(i+k)}(x_1) \right| && (= R_{f^{(i)}}(x_1, x_2 - x_1)) \\
&\qquad + \left| \sum_{k=0}^{\beta-i} \frac{(x_2 - x_1)^k}{k!} f^{(i+k)}(x_1) - \sum_{k=0}^{\beta-i} \frac{(x_2 - x_1)^k}{k!} \epsilon^{1-(i+k)/\alpha} \cdot \left\lfloor \frac{f^{(i+k)}(x_1)}{\epsilon^{1-(i+k)/\alpha}} \right\rfloor \right| \\
&\le \frac{(x_2 - x_1)^{\alpha-i}}{(\beta - i)!} + \sum_{k=0}^{\beta-i} \frac{(x_2 - x_1)^k}{k!} \left| \epsilon^{1-(i+k)/\alpha} \left\lfloor \frac{f^{(i+k)}(x_1)}{\epsilon^{1-(i+k)/\alpha}} \right\rfloor - f^{(i+k)}(x_1) \right| \\
&\le \frac{\epsilon^{1-i/\alpha}}{(\beta - i)!} + \sum_{k=0}^{\beta-i} \frac{\epsilon^{k/\alpha}}{k!} \epsilon^{1-(i+k)/\alpha} \left| \left\lfloor \frac{f^{(i+k)}(x_1)}{\epsilon^{1-(i+k)/\alpha}} \right\rfloor - \frac{f^{(i+k)}(x_1)}{\epsilon^{1-(i+k)/\alpha}} \right| \\
&\le \frac{\epsilon^{1-i/\alpha}}{(\beta - i)!} + \sum_{k=0}^{\beta-i} \frac{\epsilon^{1-i/\alpha}}{k!} \\
&\le \frac{\epsilon^{1-i/\alpha}}{(\beta - i)!} + \epsilon^{1-i/\alpha} e && e = \textstyle\sum_k 1/k! \\
&=: C(\alpha)\epsilon^{1-i/\alpha}.
\end{aligned}$$

⊛

[a] In the problem statement, a factor of $\epsilon^{1-(i+k)/\alpha}$ is missing.

**Claim.** If a function $f \in \mathcal{S}_\alpha$ is constrained such that $\lfloor f^{(k)}(x_1)/\epsilon^{1-k/\alpha} \rfloor = A_k$ for $k = 0, \dots, \beta$ for some integers $A_k$, $k = 0, \dots, \beta$, the number of possible values of $\lfloor f^{(k)}(x_2)/\epsilon^{1-k/\alpha} \rfloor = A_k$ for $k = 0, \dots, \beta$ is at most a constant $C(\alpha)$ depending only on $\alpha$. The same conclusion holds if $x_1$ and $x_2$ are replaced by $x_j$ and $x_{j+1}$ for every $1 \le j \le s - 1$.

**Proof.** Consider a function $f \in \mathcal{S}_\alpha$ with constraints at $x_1$ such that $\left\lfloor \frac{f^{(k)}(x_1)}{\epsilon^{1-(k/\alpha)}} \right\rfloor = A_k$ for some integers $A_k$, $k = 0, \dots, \beta$. To bound the number of possible values of $\left\lfloor \frac{f^{(k)}(x_2)}{\epsilon^{1-k/\alpha}} \right\rfloor$ for $k = 0, \dots, \beta$, note that by relabelling the previous claim, for all $k = 0, \dots, \beta$, we have

$$\left| f^{(k)}(x_2) - \sum_{\ell=0}^{\beta-k} \frac{(x_2 - x_1)^\ell}{\ell!} \epsilon^{1-(k+\ell)/\alpha} A_{k+\ell} \right| \le C(\alpha)\epsilon^{1-(k/\alpha)},$$

hence

$$\left| \frac{f^{(k)}(x_2)}{\epsilon^{1-k/\alpha}} - \sum_{\ell=0}^{\beta-k} \frac{(x_2 - x_1)^\ell}{\ell!} \epsilon^{1-(k+\ell)/\alpha} \frac{A_{k+\ell}}{\epsilon^{1-(k/\alpha)}} \right| \le C(\alpha) \Leftrightarrow \left| \frac{f^{(k)}(x_2)}{\epsilon^{1-k/\alpha}} - \sum_{\ell=0}^{\beta-k} \frac{A_{k+\ell}}{\ell!} \right| \le C(\alpha)$$

since $(x_2 - x_1)^\ell = \epsilon^{\ell/\alpha}$.

This implies that for a fixed $k$, with the fact that $A_{k+\ell}$'s are all fixed, $\sum_\ell A_{k+\ell}/\ell!$ is fixed as well, i.e., the possible values of $f^{(k)}(x_2)/\epsilon^{1-k/\alpha}$ spreads among an interval with width $2C(\alpha)$ centered at $\sum_\ell A_{k+\ell}/\ell!$. In all, up to some constants, the number of possible values of $\left\lfloor \frac{f^{(k)}(x_2)}{\epsilon^{1-k/\alpha}} \right\rfloor$ for $k = 0, \dots, \beta$ is at most a constant $C(\alpha)$ depending only on $\alpha$. Finally, for general $j$, as $x_{j+1} - x_j = \epsilon^{1/\alpha}$ for $1 \le j \le s-2$, and $\le \epsilon^{1/\alpha}$ when $j = s-1$, we see that the inequalities all hold, hence the same conclusion can be reached for all $(x_j, x_{j+1})$ for $1 \le j \le s-1$. ⊛

We can finally show the upper-bound.

**Claim.** We have

$$|I| \le \left(\frac{1}{\epsilon}\right)^{\beta+1} (C(\alpha))^s \le \exp\left(C(\alpha)\epsilon^{-1/\alpha}\right).$$

**Proof.** Observe that $|I|$ can be bounded by first consider $\left\lfloor \frac{f^{(k)}(x_1)}{\epsilon^{1-(k/\alpha)}} \right\rfloor$, which has $C(\alpha)\epsilon^{-\beta-1}$ possibilities from the previous claim (for every $k$ and $f \in \mathcal{S}_\alpha$). Then for every consecutive $x_{i+1}$, it introduces $C(\alpha)$ more possibilities for the tuple

$$\left( \left\lfloor \frac{f^{(k)}(x_i)}{\epsilon^{1-(k/\alpha)}} \right\rfloor \right)_{\substack{i=1,\dots,s \\ k=0,\dots,\beta}},$$

hence in total,

$$|I| \le C(\alpha)\epsilon^{-\beta-1} \cdot C(\alpha)^{s-1} = \left(\frac{1}{\epsilon}\right)^{\beta+1} (C(\alpha))^s$$

We further see that since $s \approx \epsilon^{-1/\alpha}$ and $\ln 1/\epsilon \le C(1/\epsilon)^{1/\alpha}$ for some constant $C$,

$$\ln |I| \le (\beta+1)\ln 1/\epsilon + s \ln C(\alpha) \le C(\alpha) \cdot (1/\epsilon)^{1/\alpha} + \epsilon^{-1/\alpha} \ln C(\alpha) \le C(\alpha)\epsilon^{-1/\alpha}.$$

Hence, in all, we have

$$|I| \le \exp\left(C(\alpha)\epsilon^{-1/\alpha}\right).$$

⊛

In all, from the first claim, $N(\mathcal{S}_\alpha, d, C(\alpha)\epsilon) \le |I| \le \exp\left(C(\alpha)\epsilon^{-1/\alpha}\right)$ as desired, hence

$$M(\mathcal{S}_\alpha, d, \epsilon) \le \exp\left(c_1\epsilon^{-1/\alpha}\right)$$

for some $c_1$. As for the lower-bound, consider the following. Check

**Claim.** There exists a function $f_0 \colon \mathbb{R} \to \mathbb{R}$ such that

(a) $f_0(x) = 0$ for $x \notin (0,1)$;

(b) $f_0(x) > 0$ for $x \in (0,1)$;

(c) $f_0$ restricted to the interval $[0,1]$ lies in $\mathcal{S}_\alpha$.

**Proof.** Consider the function

$$f_0(x) = c \cdot e^{-1/x} e^{-1/(1-x)} \cdot \mathbb{1}_{0<x<1}$$

where $c > 0$ is not determined yet. We see that

(a) for $x \notin (0,1)$, $f_0(x) = 0$ as $\mathbb{1}_{0<x<1} = 0$ for $x \notin (0,1)$;

(b) for $x \in (0,1)$, $f_0(x) = c \cdot \exp(-(1/x + 1/(1-x))) = ce^{\frac{1}{x(x-1)}} > 0$ as $\exp(\cdot) > 0$ always;

(c) to show $f_0|_{[0,1]} = ce^{\frac{1}{x(x-1)}} \in \mathcal{S}_\alpha$:

- $f_0$ is continuous on $[0,1]$: true since as $x$ tends to singularities, e.g., $x \to 0$ or $x \to 1$, $f_0(x) \to c$;
- $f$ is $\beta$-times differentiable: true trivially;
- $|f^{(k)}| \le 1$ for all $k = 0, \dots, \beta$: we see that as $f$ is smooth (only singularities are at the boundary, i.e., 0 and 1), there is no $k$ such that $|f^{(k)}(x)| = \infty$ for $x \in (0,1)$, so by choosing an appropriate $c$ to re-normalize, $|f^{(k)}| \le 1$ for all $k$;
- $|f^{(\beta)}(x) - f^{(\beta)}(y)| \le |x-y|^{\alpha-\beta}$ for all $x, y \in [0,1]$: as $f_0$ is bounded on $[0,1]$ (hence Lipschitz), $|f'|$ is bounded as well, which in itself is Lipschitz, etc. By iteratively using this argument, $f^{(\beta)}$ is Lipschitz as well.[a]

⊛

[a] Note that we need to rescale to make sure it satisfies the exact bound.

**Claim.** Consider points $0 < a_1 < b_1 < a_2 < b_2 < \cdots < a_s < b_s < 1$ where $b_i - a_i = \epsilon^{1/\alpha}$ and $s \ge C(\alpha)\epsilon^{-1/\alpha}$. For each $i = 1, \dots, s$, define

$$g_i(x) \coloneqq (b_i - a_i)^\alpha f_0\left(\frac{t - a_i}{b_i - a_i}\right)$$

where $f_0$ is as in the previous claim. Then $g_i \in \mathcal{S}_\alpha$ and that $g_i$ is supported on $(a_i, b_i)$.

**Proof.** Firstly, $g_i$ defined as

$$g_i(x) \coloneqq (b_i - a_i)^\alpha f_0\left(\frac{x - a_i}{b_i - a_i}\right)$$

is clearly supported on $(a_i, b_i)$ since $f_0$ has a support on $(0,1)$ so $f_0((x-a_i)/(b-i-a_i))$ has a support on $(a_i, b_i)$. Since now $g_i(x)$ is a "squeezed" version of $f_0|_{[0,1]}$, the re-normalized factor $(b_i - a_i)^\alpha$ brings the height down to make it in $\mathcal{S}_\alpha$. Specifically,

- $g_i$ is clearly continuous;
- $g_i$ is still $\beta$-times differentiable;
- $|g_i^{(k)}| \le 1$ for all $k = 0, \dots, \beta$: we see that

  $$g_i^{(k)}(x) = (b_i - a_i)^{\alpha-k} f_0^{(k)}\left(\frac{x - a_i}{b_i - a_i}\right);$$

  and since $|f^{(k)}| \le 1$ and $(b_i - a_i)^{\alpha-k} \le 1$ as well, $|g_i^{(k)}| \le 1$;
- $|g_i^{(\beta)}(x) - g_i^{(\beta)}(y)| \le |x - y|^{\alpha-\beta}$ for all $x, y \in [0,1]$: since $f_0 \in \mathcal{S}_\alpha$,

  $$\begin{aligned} |g_i^{(\beta)}(x) - g_i^{(\beta)}(y)| &= (b_i - a_i)^{\alpha-\beta} \left| f_0^{(\beta)}\left(\frac{x - a_i}{b_i - a_i}\right) - f_0^{(\beta)}\left(\frac{y - a_i}{b_i - a_i}\right)\right| \\ &\le (b_i - a_i)^{\alpha-\beta} \cdot \left|\frac{x - y}{b_i - a_i}\right|^{\alpha-\beta} \\ &\le |x - y|^{\alpha-\beta}. \end{aligned}$$

Hence, $g_i \in \mathcal{S}_\alpha$ as well. ⊛

**Claim.** For every $\tau \in \{0,1\}^s$, define $u_\tau(x) \coloneqq \sum_{i=1}^s \tau_i g_i(x)$. Then $u_\tau \in \mathcal{S}_\alpha$ for every $\tau \in \{0,1\}^s$.

**Proof.** For every $\tau \in \{0,1\}^s$, consider $u_\tau(x) = \sum_{i=1}^s \tau_i g_i(x)$. We see that

- $u_\tau(x)$ is continuous: since $g_i$'s boundary coincides with $f_0$'s boundary, which has value 0, so concatenate them together is continuous;
- $u_\tau(x)$ is $\beta$-times differentiable: as $f_0$ is symmetry around $1/2$, $g_i$ is now symmetry around $(b_i - a_i)/2$; moreover, since the higher-order differentiation of $f_0$ at the boundaries vanishes to 0, $u_\tau(x)$ is actually $\beta$-times differentiable;
- $|u_\tau^{(k)}| \le 1$ for all $k = 0, \dots, \beta$: since $g_i^{(k)}$'s have separate supports,

  $$|u_\tau^{(k)}| = \left|\sum_{i=1}^s \tau_i g_i^{(k)}(x)\right| \le \max_i |g_i^{(k)}(x)| \le 1;$$

- $|u_\tau^{(\beta)}(x) - u_\tau^{(\beta)}(y)| \le |x - y|^{\alpha-\beta}$: same reason as above.

⊛

**Claim.** For every $\tau, \tau' \in \{0,1\}^s$ with $\tau_j \ne \tau'_j$ for some $j$, $d(u_\tau, u'_\tau) \ge f_0(1/2)\epsilon$.

**Proof.** Consider two $\tau, \tau' \in \{0,1\}^s$ such that $\tau_j \neq \tau'_j$ for some $j$, we see that

$$d(u_\tau, u_{\tau'}) = \sup_{0<x<1} \left| \sum_{i=1}^{s} \tau_i g_i(x) - \sum_{i=1}^{s} \tau'_i g_i(x) \right| = \sup_{0<x<1} \left| \sum_{j:\ \tau_j \neq \tau'_j} g_j(x) \right| = \sup_{a_j<x<b_j} |g_i(x)|$$

for some specific $j$ such that $\tau_j \neq \tau'_j$.[a] Now, we have

$$d(u_\tau, u_{\tau'}) = \sup_{a_j<x<b_j} |g_j(x)| = \sup_{a_j<x<b_j} (b_j - a_j)^\alpha f_0\left(\frac{x-a_j}{b_j-a_j}\right) \geq \epsilon f_0(1/2)$$

from $b_j - a_j = \epsilon^{1/\alpha}$ and by choosing some arbitrary $(x-a_j)/(b_j-a_j)$ (in this case, we just choose $1/2$). ⊛

[a] This holds since $g_i$'s are disjoint.

We conclude that the set $\{u_\tau \in \mathcal{S}_\alpha\}_{\tau\in\{0,1\}^s}$ is not an $f_0(1/2)\epsilon$-net, i.e.,

$$\left|\{u_\tau \in \mathcal{S}_\alpha\}_{\tau\in\{0,1\}^s}\right| \leq N(\mathcal{S}_\alpha, d, f(1/2)\epsilon).$$

With the fact that

$$\left|\{u_\tau \in \mathcal{S}_\alpha\}_{\tau\in\{0,1\}^s}\right| = 2^s \geq 2 \cdot C(\alpha)\epsilon^{-1/\alpha} =: C(\alpha)\epsilon^{-1/\alpha},$$

we finally have

$$C(\alpha)\epsilon^{-1/\alpha} \leq \left|\{u_\tau \in \mathcal{S}_\alpha\}_{\tau\in\{0,1\}^s}\right| \leq N(\mathcal{S}_\alpha, d, f(1/2)\epsilon).$$

■

We now provide some missing proofs for different forms of Dudley's entropy bound.

**Corollary A.2.1** (High probability form (Corollary 3.3.3)). The high probability bound version holds:

$$\mathbb{P}\left(\sup_{s,t\in T} |X_s - X_t| \leq C\left(\int_0^\infty \sqrt{\log N(T,d,\epsilon)}\,\mathrm{d}\epsilon + u\,\mathrm{diam}(T)\right)\right) \geq 1 - 2e^{-u^2}.$$

**Proof.** Adopting the same notation as in the proof of Dudley's entropy bound for $K_0$, $K_1$, $N_k$, and $\pi_k(t)$. By writing

$$\begin{aligned} X_t - X_{t_0} &= X_{\pi_{K_1}(t)} - X_{\pi_{K_0}(t)} \\ &= X_{\pi_{K_1}(t)} - X_{\pi_{K_1-1}(t)} + X_{\pi_{K_1-1}(t)} - \cdots + X_{\pi_{K_0+1}(t)} - X_{\pi_{K_0}(t)} \\ &= \sum_{k=K_0+1}^{K_1} X_{\pi_k(t)} - X_{\pi_{k-1}(t)}, \end{aligned}$$

which implies

$$\sup_{t\in T} X_t - X_{t_0} \leq \sum_{k=K_0+1}^{K_1} \sup_{t\in T} \left(X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\right).$$

To prove a tail bound, we claim the following.

**Claim.** If $\{X_t\}_{t\in T}$ is $\mathrm{Subg}(\sigma^2)$, for all $u \geq 0$,

$$\Pr\left(\sup_{t\in T} X_t \geq \sqrt{2\sigma^2 \log|T|} + u\right) \leq \exp\left(-\frac{u^2}{2\sigma^2}\right)$$

**Proof.** From Chernoff bound,

$$\Pr\left(\sup_{t\in T} X_t \geq u\right) = \Pr\left(\bigcup_{t\in T}\{X_t \geq u\}\right) \leq \sum_{t\in T}\Pr(X_t \geq u) \leq |T|e^{-u^2/2\sigma^2}.$$

Now, let $u' \coloneqq \sqrt{2\sigma^2 \log|T|} + u$, we obtain

$$\Pr\left(\sup_{t\in T} X_t \geq u'\right) \leq \exp\left(\log|T| - \frac{2\sigma^2\log|T| + 2\sqrt{2\sigma^2\log|T|}u + u^2}{2\sigma^2}\right) \leq \exp\left(-\frac{u^2}{2\sigma^2}\right).$$

⊛

Then, since $X_{\pi_k(t)} - X_{\pi_{k-1}(t)} \sim \mathrm{Subg}(d^2(\pi_k(t), \pi_{k-1}(t)))$ with

$$d(\pi_k(t), \pi_{k-1}(t)) \leq d(\pi_k(t), t) + d(t, \pi_{k-1}(t)) \leq 2^{-k} + 2^{-k+1} \leq 3\cdot 2^{-k}$$

from the above claim,

$$\Pr\left(\sup_{t\in T} X_{\pi_k(t)} - X_{\pi_{k-1}(t)} \geq 6\times 2^{-k}\sqrt{\log|N_k|} + 3\times 2^{-k}z\right) \leq e^{-z^2/2}$$

by letting $u = 3\cdot 2^{-k}z$. Now, we apply a union bound over $k$ with $z_k \coloneqq u + \sqrt{k-K_0}$, which yields

$$\begin{aligned}
&\Pr\left(\exists k\colon \sup_{t\in T} X_{\pi_k(t)} - X_{\pi_{k-1}(t)} \geq 6\times 2^{-k}\sqrt{\log|N_k|} + 3\times 2^{-k}z_k\right)\\
&\leq \sum_{k=K_0+1}^{K_1}\Pr\left(\sup_{t\in T} X_{\pi_k(t)} - X_{\pi_{k-1}(t)} \geq 6\times 2^{-k}\sqrt{\log|N_k|} + 3\times 2^{-k}z_k\right)\\
&\leq \sum_{k=K_0+1}^{K_1} e^{-z_k^2/2}\\
&= \sum_{k=K_0+1}^{K_1}\exp\left(-\frac{u^2 + 2x\sqrt{k-K_0} + (k-K_0)}{2}\right)\\
&\leq e^{-u^2/2}\sum_{k=1}^{\infty} e^{-k/2}\\
&\leq 2e^{-u^2/2}. \qquad \textstyle\sum_k e^{-k/2} = \frac{1}{e^{1/2}-1} \approx 1.541
\end{aligned}$$

This means that with probability at least $1 - 2e^{-u^2/2}$,

$$\begin{aligned}
\sup_{t\in T} X_t - X_{t_0} &\leq \sum_{k=K_0+1}^{K_1}\sup_{t\in T}\left(X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\right)\\
&\leq \sum_{k=K_0+1}^{K_1} 6\times 2^{-k}\sqrt{\log|N_k|} + 3\times 2^{-k}z_k\\
&\leq 6\sum_{k>K_0} 2^{-k}\sqrt{\log|N_k|} + 3\times 2^{-K_0}\sum_{k>0} 2^{-k}\sqrt{k} + 3\times 2^{-K_0}\sum_{k>0} 2^{-k}u\\
&\leq C\left(\int_0^{\infty}\sqrt{\log N(T,d,\epsilon)}\,\mathrm{d}\epsilon + u\,\mathrm{diam}(T)\right)
\end{aligned}$$

since $2^{-K_0} \leq 2\,\mathrm{diam}(T)$ and

$$2^{-K_0} \leq C2^{-K_0-1}\sqrt{\log N(T,d,2^{-K_0-1})} \leq C\sum_{k>K_0} 2^{-k}\sqrt{\log|N_k|}.$$

Finally, observe that by considering the same bound for $X_s - X_{t_0}$ and use triangle inequality, we obtain the desired result. ■

**Corollary A.2.2** (Finite resolution form (Corollary 3.3.4)). The following generalizes the Dudley's integral entropy bound in the sense that $\delta > 0$:

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \le C\left(\mathbb{E}\left[\sup_{\substack{t,t'\in T\\ d(t,t')\le\delta}} X_t - X_{t'}\right] + \int_\delta^\infty \sqrt{\log N(T,d,\epsilon)}\,\mathrm{d}\epsilon\right).$$

**Proof.** Again, we adopt the same notation as in the proof of Dudley's entropy bound for $K_0$, $K_1$, $N_k$, and $\pi_k(t)$. Moreover, for any $t, t' \in T$, let $N_{k_\delta}$ be a minimal $\delta$-net,[a] then we have

$$\begin{aligned} X_t - X_{t'} &= X_t - X_{\pi_{k_\delta}(t)} + X_{\pi_{k_\delta}(t)} - X_{\pi_{k_\delta}(t')} + X_{\pi_{k_\delta}(t')} - X_{t'} \\ &\le 2\sup_{\substack{t,t'\in T:\\ d(t,t')\le\delta}} (X_t - X_{t'}) + \sup_{\hat{t},\hat{t}'\in N_{k_\delta}} (X_{\hat{t}} - X_{\hat{t}'}). \end{aligned}$$

Hence, we have

$$\begin{aligned} \mathbb{E}\left[\sup_{t\in T} X_t\right] = \mathbb{E}\left[\sup_{t\in T} X_t - X_{t'}\right] &\le \mathbb{E}\left[\sup_{t,t'} X_t - X_{t'}\right] \\ &\le 2\mathbb{E}\left[\sup_{\substack{t,t'\in T:\\ d(t,t')\le\delta}} (X_t - X_{t'})\right] + \mathbb{E}\left[\sup_{\hat{t},\hat{t}'\in N_{k_\delta}} (X_{\hat{t}} - X_{\hat{t}'})\right]. \end{aligned}$$

We can then handle the second term as in the proof of Dudley's entropy bound, just that now we're summing over $k$ from $k_\delta + 1$, not $K_0 + 1$. With the fact that when making the chaining sum into integral turns the lower limit into $2^{-k_\delta} = \delta$, we're done. ■

[a] $k_\delta$ is determined by $\delta$; specifically, $2^{-k_\delta} = \delta$.

# Bibliography


[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013. ISBN: 978-0-19-953525-5. URL: https://books.google.com/books?id=5oo4YIz6tR0C.

[Cha14] Sourav Chatterjee. "A new perspective on least squares under convex constraint". In: *The Annals of Statistics* 42.6 (2014). DOI: 10.1214/14-aos1254. URL: https://doi.org/10.1214%2F14-aos1254.

[GRS19] Noah Golowich, Alexander Rakhlin, and Ohad Shamir. *Size-Independent Sample Complexity of Neural Networks*. 2019. arXiv: 1712.06541 [cs.LG].

[Han16] Ramon van Handel. "Probability in High Dimensions". In: (2016). URL: https://web.math.princeton.edu/~rvan/APC550.pdf.

[Nov62] Albert BJ Novikoff. "On convergence proofs on perceptrons". In: *Proceedings of the Symposium on the Mathematical Theory of Automata*. Vol. 12. 1. New York, NY. 1962, pp. 615–622.

[OH10] Samet Oymak and Babak Hassibi. *New Null Space Results and Recovery Thresholds for Matrix Rank Minimization*. 2010. arXiv: 1011.6326 [math.OC].

[OH13] Samet Oymak and Babak Hassibi. *Sharp MSE Bounds for Proximal Denoising*. 2013. arXiv: 1305.2714 [cs.IT].

[Tsy08] A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York, 2008. ISBN: 978-0-387-79052-7. URL: https://books.google.com/books?id=mwB8rUBsbqoC.

[Vaa98] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press, 1998. ISBN: 978-0-521-78450-4. DOI: 10.1017/CBO9780511802256. URL: https://www.cambridge.org/core/books/asymptotic-statistics/A3C7DAD3F7E66A1FA60E9C8FE132EE1D (visited on 10/17/2023).

[VW96] Aad W. Van Der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. New York, NY: Springer, 1996. ISBN: 978-1-4757-2547-6 978-1-4757-2545-2. DOI: 10.1007/978-1-4757-2545-2. (Visited on 08/21/2023).