

STAT576
Empirical Process Theory

Pingbang Hu

September 20, 2023

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 covering page is the last updated time.

Contents

1	Introduction	2
1.1	What is Empirical Process Theory?	2
1.2	Applications of Uniform Law of Large Numbers	3
1.3	Bounding Supremum of Empirical Process	5
2	Concentration Inequalities	6
2.1	Gaussian Distribution	6
2.2	MGF Trick	7
2.3	Hoeffding's Inequality	8
2.4	Bernstein's Inequality	11
2.5	Bounded Difference Concentration Inequality	14
3	Expected Supremum of Empirical Process	20
3.1	Goodness of Fit Testing	20
3.2	Statistical Learning	20
3.3	Glivenko-Cantelli Class and Vapnik-Chervonenkis Dimension	26
3.4	Metric Entropy Methods	30

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

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

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

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

On the other hand, we can also ask does this convergence happen if we jointly consider all possible $t \in \mathbb{R}$. By the [Glivenko-Cantelli theorem](#), $\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \xrightarrow{n \rightarrow \infty} 0$ 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 χ be the domain, \mathbb{P} be a distribution on χ , and \mathcal{F} be the class of function such that $\chi \rightarrow \mathbb{R}$. The *empirical process* is the stochastic process indexed by functions in \mathcal{F} , $\{G_n(f) : f \in \mathcal{F}\}$ where

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

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

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 arises.

1.1.1 Uniform Law of Large Numbers

As $n \rightarrow \infty$, whether

$$S_n(\mathcal{F}) := \sup_{f \in \mathcal{F}} |G_n(f)| \rightarrow 0,$$

and if, at what rate?

Remark. The rate of convergence of law of large numbers uniformly over a class of functions \mathcal{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 $S(\mathcal{F})$ concentrates around its expectation and will bound $\mathbb{E}[S(\mathcal{F})]$.

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, $\sqrt{n}G_n(f) \xrightarrow{d} \mathcal{N}(0, \text{Var}[f(X)])$ for any f . We want to understand if this holds uniformly (jointly) over $f \in \mathcal{F}$ in some sense. \circledast

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
- $\mathcal{F} = \{\mathbb{1}_{[-\infty, t]} : t \in \mathbb{R}\}$
- $U_n(t) = \sqrt{n}(F_n(t) - t)$ where F_n is the [empirical CDF](#).

We can view $U_n(t)$ as 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 $t \in [0, 1]$, $U_n(t) \rightarrow \mathcal{N}(0, t - t^2)$ as $n \rightarrow \infty$. Moreover, for fixed t_1, \dots, t_k , the multivariate CLT implies that $(U_n(t_1), \dots, U_n(t_k)) \xrightarrow{d} \mathcal{N}(0, \Sigma)$ where $\Sigma_{ij} = \min(t_i, t_j) - t_i t_j$.

^a \mathcal{U} denotes the uniform distribution.

From this example, one can ask question 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. \circledast

Other than that, how do we characterize convergence of stochastic processes in distribution to another stochastic process? How do we generalize this result for a general function class \mathcal{F} defined on a probability space χ ? 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 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 χ , Θ is the parameter space, and $M_{\theta}: \chi \rightarrow \mathbb{R}$ 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 $\{p_{\theta}: \theta \in \Theta\}$, then $\hat{\theta}$ is the *Maximum log-likelihood* of θ .

There are lots of examples on “local estimators” as well.

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

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

Example (τ 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

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

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

Problem. Is $\hat{\theta}$ consistent for θ_0 ? Does $\hat{\theta}$ converge to θ_0 almost surely or in probability as $n \rightarrow \infty$? I.e., is $d(\hat{\theta}, \theta_0) \rightarrow 0$ 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 $M(\theta) = \mathbb{E} [M_{\theta}(X_1)]$, then

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

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

1.3 Bounding Supremum of Empirical Process

Most of this course will focus on bounding suprema of the [empirical process](#). Let's define it rigorously.

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

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

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

- (a) $S_n(\mathcal{F})$ “concentrates” around its expectation $\mathbb{E}[S_n(\mathcal{F})]$.
- (b) $\mathbb{E}[S_n(\mathcal{F})] \leq$ the [Rademacher complexity](#) of \mathcal{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 which 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 on concentration inequalities. The most celebrated concentration inequality might be the Gaussian tail, which achieves a quadratic exponential decay. Combine this with the classical central limit theorem, we can expect that as $n \rightarrow \infty$, approximately the Gaussian tail bound kicks in.

However, to get a concrete, non-asymptotic bound for $S_n(\mathcal{F})$, 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 $Z \sim \mathcal{N}(0, 1)$,

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

Proof. We want to show

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

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

$$\int_t^\infty \frac{x}{x} \cdot e^{-x^2/2} dx = -\frac{e^{-x^2/2}}{x} \Big|_t^\infty - \int_t^\infty \frac{e^{-x^2/2}}{x^2} dx = \frac{e^{-t^2/2}}{t} - \int_t^\infty \frac{e^{-x^2/2}}{x^2} dx \leq \frac{1}{t} \cdot e^{-t^2/2}$$

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

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

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, \dots, X_n with mean μ and variance σ^2 , whether for all $t \geq 0$,

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

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 Gaussian tail bound, let start with the [moment generating function \(MGF\) trick](#).

2.2 MGF Trick

The [MGF trick](#) is easy to develop, but it gives a foundation of 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 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

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

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 Cramer-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 (Cramer-Chernoff method)). Given a random variable X ,

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

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^{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 σ^2 if for all $\lambda \in \mathbb{R}$,

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

^aAlso called proxy, sub-Gaussian norm, etc.

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

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

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

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

- (a) $\mathbb{E}[e^{\lambda X}] \leq e^{c_1^2 \lambda^2}$ for all $\lambda \in \mathbb{R}$.
- (b) $\mathbb{P}(|X| \geq t) \leq 2e^{-t^2/c_2^2}$.
- (c) $(\mathbb{E}[|X|^p])^{1/p} \leq c_3 \sqrt{p}$.

Add proof

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

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

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

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

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. ■

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

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

since $(2k)! \geq 2^k \cdot k!$. *

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

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

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 $X' \stackrel{\text{i.i.d.}}{\sim} X$ be an i.i.d. copy, then

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

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

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

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

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

■

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

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

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

■

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

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

^aOne-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 μ_i ,

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

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

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

Compare this with Gaussian approximation, we then have

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

i.e., $\sigma^2 \sim (b-a)^2/4$.¹

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 $\sigma^2 \leq (b-a)^2/4$, 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

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

for μ as

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

while from the [Hoeffding's inequality](#), we have (finite sample valid) confidence interval

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

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 $\text{Var}[X] \leq 1$. 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(λ/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

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

This suggests that $\sum_i X_i = O(\sqrt{n})$ whereas we know that in this case that the distribution of $\sum_i X_i$ is close to the Poisson(λ) 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 by 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 which are not [sub-Gaussians](#) random variables are Cauchy, exponential, and Poisson random variables.

What about mixture?

Problem. Suppose $Z_1, Z_2 \in \text{Subg}(\sigma^2)$ 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 fat 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. \circledast

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 χ^2 tail probability is slower than that of normal. In fact, the MGF of χ^2 is

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

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 (σ^2, α) with mean λ 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 λ in the interval $(-\frac{1}{\alpha}, \frac{1}{\alpha})$.

Example. For the χ^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

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

\circledast

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 μ . Then

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

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 α :

- $\frac{t}{\sigma^2} < \frac{1}{\alpha}$: 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 μ . Then

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

for all $t \geq 0$.

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 μ_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 μ_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 μ_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

2.5 Bounded Difference Concentration Inequality

28 Aug. 9:00

2.5.1 Applications of Bernstein'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 $|X - \mu| \leq b$ 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 $(X - \mu)^k \leq (X - \mu)^2 |X - \mu|^{k-2} \leq (X - \mu)^2 b^{k-2}$, we have

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

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

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

■

From this, by directly apply [Bernstein's inequality](#), we have the following.

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

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

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 $|X_i - \mu_i| \leq b$ for all i . Then for any $t > 0$,

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

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

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

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

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

In particular,

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

From [Corollary 2.5.2](#):

- if $t \leq 3\sigma^2/b$, 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 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 μ .

Corollary 2.5.3. Under the assumption of [Corollary 2.5.2](#),

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

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 the 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 large range but highly concentrated around its mean, the [Hoeffding confidence interval](#) would be much wider.

The above remark is demonstrated by the following example.

Example. Let $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(p)$. Suppose we observe $X_i = 0$ for all i , then $\hat{p} = \bar{X} = 0$ 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

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

“concentrates” when \mathcal{F} is bounded provided that

$$\sup_{x \in \mathcal{X}, f \in \mathcal{F}} |f(x)| \leq B.$$

One simple example of 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](#).²

Theorem 2.5.1 (McDiarmid's inequality). Let X_1, \dots, X_n be i.i.d. random variables on χ , and let $f: \chi^n \rightarrow \mathbb{R}$ satisfying the *bounded difference property*, i.e.,

$$\sup_{x_1, \dots, x_n, x'_i} |f(x_1, \dots, x_n) - f(x_1, \dots, x'_i, \dots, x_n)| \leq c_i$$

for all i . Then for any $t > 0$,

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

The same bound holds for the left tail.

Remark. The qualitative statement for [McDiarmid's inequality](#) is that “a random variable that 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. \circledast

Remark. [McDiarmid's inequality](#) is a generalization of [Hoeffding's inequality](#).

Proof. Let

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

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^{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 $c_i := (b_i - a_i)/n$, plugging in, we get back [Hoeffding's inequality](#). \circledast

With [McDiarmid's inequality](#), we can check that the following holds for bounded function classes \mathcal{F} :

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

Then from [McDiarmid's inequality](#), for any $t > 0$,

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

or equivalently,

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

with probability at least $1 - \delta$.

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

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

²It's also known as the *bounded difference inequality*.

⊛

^aThis 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

We should note that the usual proof of [McDiarmid inequality](#) involves [martingale decomposition](#) and [Azuma-Hoeffding inequality](#), a generalization of [Hoeffding's inequality](#) for [martingale difference sequence](#).

1 Sep. 9:00

Definition 2.5.1 (Martingale difference sequence). A *martingale difference sequence* is a sequence of random variables Δ_1, \dots such that $\mathbb{E}[\Delta_i \mid \Delta_{i-1}] = 0$ for all i .

However, we will not go with this route; instead, we prove something weaker but trickier.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

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

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

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

■

We're now ready.

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

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

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

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

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

Notation. $\mathbb{X}^{(i)} := (X'_1, \dots, X'_i, X_{i+1}, \dots, X_n)$ and $\mathbb{X}^{[i]} := (X_1, \dots, X_{i-1}, X'_i, X_{i+1}, \dots, X_n)$.

³In 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.

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

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

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

$$\begin{aligned} &= \mathbb{E} \left[\frac{1}{2} \sum_{i=1}^n \left(f(\mathbb{X}^{(i-1)}) - f(\mathbb{X}^{(i)}) \right) \cdot \left(e^{\lambda f(\mathbb{X})} - e^{\lambda f(\mathbb{X}^{[i]})} \right) \right] \\ &\leq \mathbb{E} \left[\frac{\lambda}{2} \sum_{i=1}^n \left| f(\mathbb{X}^{(i-1)}) - f(\mathbb{X}^{(i)}) \right| \cdot \left| f(\mathbb{X}) - f(\mathbb{X}^{[i]}) \right| \cdot \left(\frac{e^{\lambda f(\mathbb{X})} + e^{\lambda f(\mathbb{X}^{[i]})}}{2} \right) \right] \\ &\hspace{15em} \text{from Lemma 2.5.2} \\ &\leq \frac{\lambda}{2} \left(\sum_{i=1}^n c_i^2 \right) \cdot M(\lambda). \end{aligned}$$

■

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

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

we only show

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

⊛

2.5.3 Applications of McDiarmid's Inequality

U-Statistics

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

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

Here're some examples of g .

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

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

Example (Wilcoxon's ranksum 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

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

implying

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

from [McDiarmid's inequality](#) with $c_i := 2B$.

Beyond McDiarmid's Inequality

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

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

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

Theorem 2.5.2 (Efron-Stein inequality). Let X_1, \dots, X_n be independent random variables, and X'_1, \dots, X'_n be i.i.d. copies of X_i 's. Then

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

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

Now, recall that by using [McDiarmid's inequality](#), we can show that for $\mathcal{F} \ni f$ 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 $\text{Var}[f(X)]$ is small, but the maximum spread (B) is very large? In this case, we would want to replace B in the inequality by $\text{Var}[f(X)]$.

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 \mathcal{F} is B -bounded, and $S_n = \sup_{f \in \mathcal{F}} |\mathbb{P}_n f - \mathbb{P}f|$. Then

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

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

3.1 Goodness of Fit Testing

6 Sep. 9:00

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

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**, $D_n \rightarrow 0$ 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) $\sqrt{n}D_n \rightarrow \sup_{t \in [0,1]} |B(t)|$ 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.

3.2 Statistical Learning

3.2.1 Empirical Risk Minimization

Consider the following problem.

Problem 3.2.1 (Empirical risk minimization). Let $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$ be n i.i.d. copies of $(X, Y) \in \mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^d \times \mathbb{R}$ with distribution $\mathbb{P} = \mathbb{P}_X \times \mathbb{P}_{Y|X}$. Given a loss function $\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ and a function class $\mathcal{F} = \{f: \mathcal{X} \rightarrow \mathcal{Y}\}$, the *empirical risk minimization* is

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

Example. \mathcal{F} can be the set of neural networks, decision trees, linear functions.

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

Example (Linear classification). Consider $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{0, 1\}$, with $\mathcal{F} = \{x \rightarrow (\text{sgn}(w^\top x) + 1)/2 : w \in B_2^d\}$ 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 3.2.1 (Expected loss). The *expected loss*^a of $f \in \mathcal{F}$ is defined as

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

^aAlso called *population loss* and *test error*.

Definition 3.2.2 (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 that, what is an upper-bound on the [expected loss](#) of [ERM](#)? If we plug in \hat{f} instead of f , this is asking the [test error](#) of \hat{f} .

To be specific, \hat{f} is basically a function of training data S , but when we look at

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

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

Lemma 3.2.1. For any \mathcal{F} , the [ERM](#) \hat{f} satisfies

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

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

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

■

Note. Let us decode what [Lemma 3.2.1](#) is claiming.

- Since $L(f)$ is the [population error](#) of f and $\hat{L}(f)$ is the [empirical loss](#) of f , $\sup_{f \in \mathcal{F}} (L(f) - \hat{L}(f))$ 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](#).

^aOr the best possible prediction error of \mathcal{F} .

Notation (Excess risk). $\mathbb{E}[L(\hat{f})] - \inf_{f \in \mathcal{F}} L(f)$ is often called the *excess risk* of an [ERM](#).

Remark. For “curved” loss function like square loss, supremum can be further “localized”.

Remark. The bound in [Lemma 3.2.1](#) 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 $b = f(x)$,

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

which can be viewed essentially as^a the [empirical process](#) on the function f instead of ℓ ,

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

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

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

i.e., $P(X \leq \theta)$ is the CDF of the marginal distribution of X , $F(\theta)$, and $\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{x_i \leq \theta}$ 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 $\mathbb{P}(X \leq \theta) - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{x_i \leq \theta}$ to be a random variable U_θ . One problem here is, 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_θ is at most 1, for any θ , i.e., $\sup U_\theta \leq 1$. So the worst case here is 1, and probably the best case is $O(1/\sqrt{n})$.

^aSince $Y - \sum_i y_i/n$ 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

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

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

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

3.2.2 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.,

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

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

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

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

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

To further bound Equation 3.1, recall the following.

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

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

^aSince it's bounded between 0 and 1.

Finally, recall what we have proved in the homework.

Lemma 3.2.2. Let $X_1, \dots, X_n \sim \text{Subg}(\sigma^2)$,^a then $\mathbb{E} [\max_i X_i] \leq \sqrt{2\sigma^2 \log n}$.

^aNot necessary independent.

Then, we can show the final bound.

Proposition 3.2.1 (Bracketing). Let $x_1, \dots, x_n \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}$, and $\mathcal{F} = \{\mathbb{1}_{X \leq \theta} : \theta \in \mathbb{R}\}$. Then

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

Proof. From Lemma 3.2.2, since we have $(N+1)$ random variables with variance factor $1/n$, by choosing $N+1 := n$,^a Equation 3.1 can be further bounded by

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

■

^aRecall that n is the sample size, so we can choose the corresponding n to meet the requirement.

3.2.3 Symmetrization

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

Lemma 3.2.3 (Symmetrization). Given a function class $\mathcal{F} = \{f: \mathcal{X} \rightarrow \mathcal{Y}\}$ and $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}$, and $\epsilon_1, \dots, \epsilon_n$ be i.i.d. [Rademacher random variables](#). Then

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

In particular,

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

Proof. Let X'_i 's be i.i.d. copies of X_i 's for all i . Since adding a sign ϵ_i won't change the expectation,^a

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

■

^aSince 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 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 3.2.3 (Rademacher complexity). Let $X_i \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}$ be independent and ϵ_i be i.i.d. [Rademacher random variables](#). The *Rademacher complexity* of a function class \mathcal{F} w.r.t. \mathbb{P} is

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

On the other hand, the opposite direction of [symmetrization lemma](#) also holds.

Lemma 3.2.4. Given a function class $\mathcal{F} = \{f: \mathcal{X} \rightarrow \mathcal{Y}\}$ and $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}$, and $\epsilon_1, \dots, \epsilon_n$ be i.i.d. [Rademacher random variables](#). Then

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

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

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

The first term can be further bounded by

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

and the second term can be bounded by

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

where $\mathbb{E} \left[\left| \frac{1}{n} \sum_{i=1}^n \epsilon_i \right| \right] \leq \frac{c}{\sqrt{n}}$ with $c = 1$. Combine them together, we have the final result. \blacksquare

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 3.2.4 (Rademacher width). Given $A \subseteq \mathbb{R}^n$, the *Rademacher width*^a of A is defined as

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

^aAlso called *Rademacher average*.

Notation. People sometimes just say “Rademacher complexity” for [Rademacher width](#).

Now, applying the [symmetrization lemma](#) to $\mathcal{F} = \{\mathbb{1}_{X \leq \theta} : \theta \in \mathbb{R}\}$, we have the following result that is comparable to [Proposition 3.2.1](#).

Proposition 3.2.2. Let $x_1, \dots, x_n \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}$, and $\mathcal{F} = \{\mathbb{1}_{x \leq \theta} : \theta \in \mathbb{R}\}$. Then

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

Proof. From the [symmetrization lemma](#),

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

Let

$$V_\theta := \frac{1}{n} \sum_i \epsilon_i \mathbb{1}_{x_i \leq \theta},$$

we see that there are only $n+1$ distinct V_θ '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 $X_{(0)} := -\infty$. Now, define $\theta_k := X_{(k)}$, we can then write

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

hence,

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

with $V_{\theta_k} \sim \text{Subg}(1/n)$ and [Lemma 3.2.2](#),

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

■

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

3.3 Glivenko-Cantelli Class and Vapnik-Chervonenkis Dimension

3.3.1 Glivenko-Cantelli Class

Definition 3.3.1 (Glivenko-Cantelli). A function class $\mathcal{F} = \{f: \chi \rightarrow \mathbb{R}\}$ is called *Glivenko-Cantelli* w.r.t. \mathbb{P} if

$$\sup_{f \in \mathcal{F}} |\mathbb{P}f - \mathbb{P}_n f| \rightarrow 0$$

as $n \rightarrow \infty$.

From [bracketing](#) and [symmetrization](#), we know that $\mathcal{F} = \{\mathbb{1}_{X \leq \theta} : \theta \in \mathbb{R}\}$ is [Glivenko-Cantelli](#). Let's see some counterexamples.

Example. Let $\chi = \mathbb{R}$, $\mathcal{F} = \{\mathbb{1}_A : A \subseteq \chi, |A| < \infty\}$, and \mathbb{P} be any continuous measure on χ . Then \mathcal{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., $\sup_{f \in \mathcal{F}} |\mathbb{P}f - \mathbb{P}_n f| = 1$ for all $n \in \mathbb{N}$. ⊗

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

Proof. Consider

- $f(X_i) = 1$ for $i = 1, \dots, n$ and $f = 0$ elsewhere (in a continuous manner).^a
- Then we can make $\int_0^1 f(t) dt < \delta$ for some $\delta \in (0, 1)$.

This implies $\sup_{f \in \mathcal{F}} |\mathbb{P}f - \mathbb{P}_n f| \geq 1 - \delta$ for all $n \in \mathbb{N}$. ⊗

^aE.g., sharp peak at X_i 's.

3.3.2 Vapnik-Chervonenkis Dimension

Let's first introduce a common notation.

Notation. Let $\mathcal{F}(x_1, \dots, x_n) := \{(f(x_1), \dots, f(x_n))\}_{f \in \mathcal{F}} \subseteq \mathbb{R}^n$.

Then we can relate the **Rademacher width** of $\mathcal{F}(X_1, \dots, X_n)$ to the **Rademacher complexity** of \mathcal{F} since

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

Note. This is why people overload these two notations.

Moreover, we see that if $\mathcal{F}(X_1, \dots, X_n)$ is finite, by the same proof as in [Proposition 3.2.2](#),

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

The up-shot is the following.

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

This is captured by the notion of **polynomial discrimination**. In particular, we'll look into the class of boolean function.

Definition 3.3.2 (Polynomial discrimination). We say that a boolean function class \mathcal{F} on χ has a *polynomial discrimination* if for all $x_1, \dots, x_n \in \chi$, $|\mathcal{F}(x_1, \dots, x_n)| \leq \text{poly}(n)$.

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

Definition. Let \mathcal{F} be a boolean function class on χ .

Definition 3.3.3 (Shatter). A finite set $\{x_1, \dots, x_D\} \subseteq \chi$ is *shattered* by \mathcal{F} if $\mathcal{F}(x_1, \dots, x_D) = \{0, 1\}^D$.

Definition 3.3.4 (Vapnik-Chervonenkis dimension). The *VC dimension* of \mathcal{F} on χ is the maximum integer D such that there exists a size D finite set $A \subseteq \chi$ *shattered* by \mathcal{F} .

Remark. We take the convention that \emptyset is always *shattered*.

Consider $\chi = \mathbb{R}$.

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

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

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

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

Lecture 9: VC Dimension

Firstly, given VC dimension, we can upper-bound the size of the discrimination.

13 Sep. 9:00

Lemma 3.3.1 (Sauer-Shelah lemma). Let \mathcal{F} be a boolean function class such that $\text{VC}(\mathcal{F}) = D$, then for every $\{x_1, \dots, x_n\} \subseteq \chi$ such that $n \geq D$,

$$|\mathcal{F}(x_1, \dots, x_n)| \leq \binom{n}{0} + \binom{n}{1} + \dots + \binom{n}{D} \leq \left(\frac{en}{D}\right)^D.$$

To prove Sauer-Shelah lemma, we need Pajor's lemma.

Lemma 3.3.2 (Pajor's lemma). Given a boolean function class \mathcal{F} on a finite set Ω , then

$$|\mathcal{F}| \leq |\{S \subseteq \Omega : S \text{ shattered by } \mathcal{F}\}|.$$

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

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

Assume the statement holds for all Ω such that $|\Omega| = n$. For $|\Omega| = n + 1$, write

$$\Omega = (\Omega \setminus \{x_0\}) \cup \{x_0\} =: \Omega_0 \cup \{x_0\}$$

and let \mathcal{F}_0 and \mathcal{F}_1 be two boolean function classes defined on Ω_0 as

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

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

$$|\mathcal{F}| = |\mathcal{F}_0| + |\mathcal{F}_1| \leq |S_{\mathcal{F}_0}| + |S_{\mathcal{F}_1}|.$$

Finally, we claim the following.

Claim. $|S_{\mathcal{F}_0}| + |S_{\mathcal{F}_1}| \leq |S_{\mathcal{F}}|$.

Proof. Let $S \subseteq \Omega_0$ shattered by both \mathcal{F}_0 and \mathcal{F}_1 , then we know S is shattered by \mathcal{F} too. We further observe that $S \cup \{x_0\}$ is shattered by \mathcal{F} , but not \mathcal{F}_i since $f(x_0)$ is fixed for $f \in \mathcal{F}_i$.

Now, when

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

By counting, the inequality always holds.^a ⊗

^aIt's possible that S is shattered by \mathcal{F} but not by \mathcal{F}_i 's, so \leq .

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

We can then prove the Sauer-Shelah lemma.

Proof of Lemma 3.3.1. Let Ω be a set of size n , then the number of subsets with size less or equal to D is

$$\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{D}.$$

By the definition of **VC dimension**, we see that

$$|\{S \subseteq \Omega: S \text{ shattered by } \mathcal{F}\}| \leq \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{D}.$$

■

Then, as our motivation suggests, the same proof of **Proposition 3.2.2** applies, giving the following.

Proposition 3.3.1. For any function class \mathcal{F} , if $n \geq \text{VC}(\mathcal{F})$, for some constant c ,

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

Remark. We see that **Proposition 3.3.1** is independent of \mathbb{P} , i.e.,

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

Remark. If $\text{VC}(\mathcal{F}) = \infty$, then “distribution-free” uniform convergence fails.

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

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

Remark. The $\sqrt{\log n}$ factors in **Proposition 3.3.1** is superfluous.

Example. Let V be a D -dimensional vector space of real function on χ , and $\mathcal{F} = \{\mathbb{1}_{f \geq 0}: f \in V\}$. Then $\text{VC}(\mathcal{F}) \leq D$.

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

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

which is a linear subspace of \mathbb{R}^{D+1} such that $\dim(T) \leq D$. This implies that 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 \mathcal{F} **shatters** $\{x_1, \dots, x_{D+1}\}$, then there exists $f \in V$ such that

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

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

Example (Half-space). Consider \mathcal{F} being the indicators of all closed half-spaces in \mathbb{R}^d . Then $\text{VC}(\mathcal{F}) = d + 1$.

It seems like the **VC dimension** is always approximately the number of parameters; however, it's not true in general.

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

Lecture 10: Discretization of a Space

3.4 Metric Entropy Methods

15 Sep. 9:00

In the previous two lectures, we focused on boolean function class with finite **VC dimension**. Now, our goal is to extend the result to functions that are not necessarily boolean. We start from studying the discretization of a space.

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

Definition 3.4.1 (Pseudo-metric). Given a space T , a function $d: T \times T \rightarrow \mathbb{R}^+$ 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) $d(x, y) \leq d(x, z) + d(y, z)$ for all $x, y, z \in T$.

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

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

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

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

3.4.1 Covering Number and Packing Number

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

Definition 3.4.2 (ϵ -net). A set N is an ϵ -net of (T, d) if for all $t \in T$, there exists $\pi(t) \in N$ such that $d(t, \pi(t)) \leq \epsilon$.

Definition 3.4.3 (Covering number). The ϵ -covering number $N(T, d, \epsilon)$ of (T, d) is defined as

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

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

Definition 3.4.4 (Totally bounded). (T, d) is *totally bounded* if for all $\epsilon > 0$, $N(T, d, \epsilon) < \infty$.

Definition 3.4.5 (ϵ -packing). A set $N \subseteq T$ is an ϵ -packing of (T, d) if for all $t \neq t'$ in N , $d(t, t') > \epsilon$.

Definition 3.4.6 (Packing number). The ϵ -packing number $M(T, d, \epsilon)$ of (T, d) is defined as

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

Lemma 3.4.1. For any $\epsilon > 0$,

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

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 2ϵ -packing and \mathcal{N} to be an ϵ -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} . \otimes

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

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

■

3.4.2 Unit balls

For simplicity, we will use the following notations.

Notation. If (T, d) and ϵ 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 3.4.1. Consider $(\mathbb{R}^d, \|\cdot\|)$ where $\|\cdot\|$ is any norm. Denote $B = \{x : \|x\| \leq 1\}$, 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

$$N \text{Vol}(\epsilon B) \geq \text{Vol}(B) \Rightarrow N\epsilon^d \geq 1.$$

With $N \leq M$ from Lemma 3.4.1, 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

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

■

Definition 3.4.7 (Metric entropy). The metric entropy of (T, d) is defined as $\log M(T, d, \epsilon)$.

Note. From Proposition 3.4.1, $\log M(\mathbb{R}^d, \|\cdot\|, \epsilon) \approx d \log 1/\epsilon$.

3.4.3 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 3.4.8 (Hölder smooth function class). Fix $\alpha > 0$, and β is the greatest integer $< \alpha$. Then the Hölder smooth function class S_α is defined to be the class of functions on $[0, 1]$ such that

- (a) f continuous on $[0, 1]$;
- (b) f is β -times differentiable;
- (c) $|f^{(k)}| \leq 1$ for all $k = 0, \dots, \beta$;
- (d) $|f^{(\beta)}(x) - f^{(\beta)}(y)| \leq |x - y|^{\alpha - \beta}$ for all $x, y \in [0, 1]$.

Note. When $\alpha = 1$, S_α 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 α increases.

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

Theorem 3.4.1. There exists c_1, c_2 such that for all $\epsilon > 0$,

$$\exp\left(c_2 \epsilon^{-1/\alpha}\right) \leq M(S_\alpha, d, \epsilon) \leq \exp\left(c_1 \epsilon^{-1/\alpha}\right).$$

Proof. 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 ϵ , resulting in $1/\epsilon$ small intervals on both the x -axis and the y -axis.

Take any function $f \in \mathcal{F}$. We construct a piece-wise constant function \tilde{f} which approximates f . On each small interval in the x -axis, we can define \tilde{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. $\sup_{x \in [0, 1]} |f(x) - \tilde{f}(x)| \leq C\epsilon$.

Proof. Since f cannot vary by more than ϵ in any interval of length ϵ . ⊗

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

Problem. What is the number of distinct \tilde{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 \tilde{f} is $\frac{1}{\epsilon}$. However, after that, in the next interval, the value of \tilde{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 \tilde{f} as

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

■

Remark. Comparing [Proposition 3.4.1](#) and [Theorem 3.4.1](#), 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

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

18 Sep. 9:00

Recall our [informal principle](#), i.e., if $\{X_t\}_{t \in T}$ is sufficiently continuous w.r.t. d , then $\mathbb{E} [\sup_{t \in T} X_t]$ is governed by metric properties (e.g., [metric entropy](#)) of T .

What we mean by “sufficiently continuous” is the following.

Definition 3.4.9 (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 , $(X_{t_1}, \dots, X_{t_k}) \sim \mathcal{N}(0, \Sigma)$.

We see that

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

where $d(t, t') = \sqrt{\mathbb{E} [(X_t - X_{t'})^2]}$.

Definition 3.4.10 (Sub-Gaussian process). A stochastic process $\{X_t\}_{t \in T}$ is a *sub-Gaussian process* w.r.t. d if $X_t - X_s \sim \text{Subg}(d^2(t, s))$.

In other words, assume $\mathbb{E} [X_t] = 0$ for all $t \in T$, and

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

for all $t \neq s \in T$.

Example (Rademacher process). Consider the [Rademacher width](#) (without normalization) 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 $X_t - X_{t'} = \langle \epsilon, t - t' \rangle \sim \text{Subg}(\|t - t'\|_2^2)$, i.e., $X_t \sim \text{Subg}$ w.r.t. $\|\cdot\|_2$.

Example (Gaussian width). The *Gaussian width* of a set $T \subseteq \mathbb{R}^n$ is defined as $X_t = \langle g, t \rangle$, i.e., we replace ϵ by g . We have $X_t \sim \text{Subg}$ w.r.t. $\|\cdot\|_2$.

Theorem 3.4.2. Denote $\text{GW}(T)$ be the [Gaussian width](#) of a set T , then

$$R_n(T) \leq \text{GW}(T) \leq \sqrt{n} R_n(T)$$

Let's look at some examples of [Rademacher width](#).

Example. $R(B_\infty^n) = n$, $R(B_2^n) = \sqrt{n}$, and $R(B_1^n) \leq \|\epsilon\|_\infty \|t\|_1 = 1$ by Holder's inequality.

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

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

hence $\{X_f\}_{f \in \mathcal{F}} \sim \text{Subg w.r.t. } \|\cdot\|_\infty$.

Definition 3.4.11 (Diameter). The *diameter* of (T, d) is defined as $\text{Diam}(T) = \sup_{t, t' \in T} d(t, t')$.

Lemma 3.4.2 (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 $\mathbb{E} [\sup_{t \in T} X_t] = \mathbb{E} [\sup_{t \in T} X_t - X_{t_0}]$ for some fixed $t_0 \in T$. Now, take an ϵ -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 3.2.2](#). By rewriting the first term, we have

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

■

Example. For [Rademacher process](#), we have $\mathbb{E} \left[\sup_{t, t' \in T: \|t - t'\| \leq \epsilon} \langle \vec{\epsilon}, t - t' \rangle \right] \leq \|\vec{\epsilon}\| \epsilon \leq \sqrt{n} \epsilon$.

Let's see some applications of [single scale bound](#).

Example. Let $T = \{(0, 0, \dots, 0), (1, 0, \dots, 0), \dots, (1, 1, \dots, 1)\} \subseteq \mathbb{R}^n$, then

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

We see that we still can't remove the $\sqrt{\log n}$ from the [single scale bound](#),

$$\sqrt{n} \epsilon + \sqrt{n} \sqrt{\log N(T, \|\cdot\|_2, \epsilon)}.$$

To remove $\log n$, one need to choose $\epsilon = O(1)$, but then $\log N(T, \|\cdot\|_2, \epsilon)$ goes to infinity, and we fail.

Example. Again, consider $X_f = \sqrt{n}(\mathbb{P}_n f - \mathbb{P} f)$ on \mathcal{F} of functions bounded by 1. Then $X_f \sim \text{Subg}(2\|f - g\|_\infty^2)$. By letting $\mathcal{F} = S_1$ ([Definition 3.4.8](#)),

$$\mathbb{E} [\sup X_f] \leq c \left(\sqrt{n} \epsilon + \sqrt{1/\epsilon} \right)$$

from [single scale bound](#) and [Theorem 3.4.1](#). By letting $\epsilon = n^{-1/3}$, this bound is minimized, giving us

$$\mathbb{E} [\sup \mathbb{P}_n f - \mathbb{P} f] \leq \frac{c}{n^{1/3}}.$$

However, this is not optimal.

Remark. The optimal bound is c/\sqrt{n} .

Lecture 12: Chaining Method

20 Sep. 9:00

Definition 3.4.12 (Separable). $X_{tt \in T}$ 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 $d(t_n, t) \rightarrow 0$ satisfying $\lim_{n \rightarrow \infty} X_{t_n} = X_t$.

It's clear that $\sup_{t \in T_0} X_t = \sup_{t \in T} X_t$.

Example. If (T, d) has a countable dense set, $\{X_t\}$ has countable sample ... almost surely, then $\{X_t\}$ is **separable**.

fix

Theorem 3.4.3 (Dudley's entropy bound). Let $\{X_t\}_{t \in T}$ be a centered, **separable sub-Gaussian process** on (T, d) w.r.t. d . Then

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

Proof. Consider the following cases.

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

Proof. Let K_0 be the largest integer such that $2^{-K_0} \geq \text{Diam}(T)$, 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 .

Recall that for $t \in T$, we write $\pi_k(t)$ for the closest element in N_k to t . In particular, $d(t, \pi_k(t)) \leq 2^{-k}$. We see that by writing

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

we have

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

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

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

where $d(\pi_k(t), t) + d(t, \pi_{k-1}(t)) \leq 2^{-k} + 2^{-(k+1)} \leq 3 \cdot 2^{-k}$, for each k ,

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

from **Lemma 3.2.2**, hence we have the result. \otimes

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

Proof. From [separability](#), there exists a countable T_0 such that $\mathbb{E} [\sup_{t \in T_0} X_t] = \mathbb{E} [\sup_{t \in T} X_t]$. Now, consider a countable approximation T_k of T_0 , we then have $\sup_{t \in T_k} X_t \rightarrow \sup_{t \in T_0} X_t$ as $k \rightarrow \infty$. Then this reduces to the finite case, with the fact that $N(T_k, d, 2^{-k}) \leq N(T, d, 2^{-k})$ for all k , we're done. \otimes

This method is called *chaining* because we're constructing a chain of $X_{\pi_k(t)}$, with smaller and smaller distance. ■

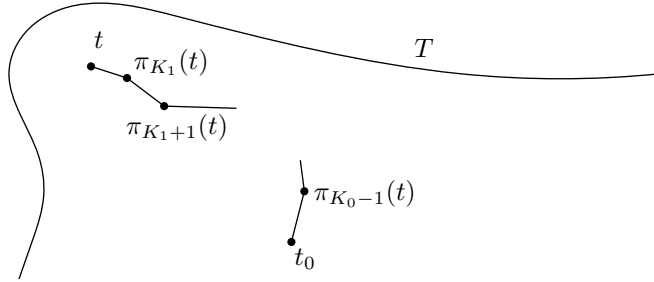


Figure 3.1: Chaining.

Remark (Dubley integral entropy bound). An alternative integral form of [Dubley entropy bound](#) is given by

$$\mathbb{E} \left[\sup_{t \in T} X_t \right] \leq 12 \int_0^{\text{Diam}(T)} \sqrt{\log N(T, d, \epsilon)} \, d\epsilon.$$

Proof. Observe that

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

\otimes

Appendix

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>.
- [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](https://doi.org/10.1007/978-1-4757-2545-2). (Visited on 08/21/2023).