STAT576
Empirical Process Theory

Pingbang Hu

September 6, 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

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, $F_n(t) \to F(t)$ 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 happen if we jointly consider all possible $t \in \mathbb{R}$. By the Glivenko-Cantelli theorem, $\sup_{t \leq \mathbb{R}} |F_n(t) - F(t)| \overset{n \to \infty}{\to} 0$ 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 χ be the domain, $\mathbb{P}$ be a distribution on χ, 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 arises.

1.1.1 Uniform Law of Large Numbers

As $n \to 0$, whether

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

and if, at what rate?

Remark. The rate of convergence of 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

- $X_1, \ldots, X_n$ i.i.d. from $\mathcal{U}(0, 1)$.[a]
- $\mathscr{F} = \left\{ \mathbb{1}_{[-\infty, t]} \colon t \in \mathbb{R} \right\}$
- $U_n(t) = \sqrt{n}(F_n(t) - t)$ 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 $\{U_n(t) \colon t \in (0, 1)\}$ 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 $t_1, \ldots, t_k$, the multivariate CLT implies that $(U_n(t_1), \ldots, U_n(t_k)) \xrightarrow{d} \mathcal{N}(0, \mathbf{\Sigma})$ 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 question like the following.

Problem. Does the entire process $\{U_n(t) \colon t \in [0, 1]\}$ 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 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 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 $X_1, \ldots, X_n$ taking values in χ, Θ 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 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

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

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

Problem. Is $\hat{\theta}$ consistent for θ_0? Does $\hat{\theta}$ converge to θ_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 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 $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, and a function class $\mathscr{F} \ni f \colon \chi \to \mathbb{R}$. We want to find an (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 Radamacher complexity 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 celebrate concentration inequality might be the Gaussian tail, which achieve a quadratic exponential decay. Combine 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 now is its rapid tail decay as we mentioned. This is given in Lemma 2.1.1.

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

Add proof

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

Proof. Omitted as a homework. ■

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 μ and variance σ^2, whether

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

for all $t \geq 0$?

Answer. Just invoking CLT is not enough, we need to handle the error term in the normal approximation. We will see that we can show the above 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 $X \geq 0$,

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

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

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

By the above remark, as 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,

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

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

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

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

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^{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 $\lambda \in \mathbb{R}$,

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

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

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

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

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

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

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

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

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

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

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 $\mathrm{Subg}(1)$ random variable.

Proof. We see that

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

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

Add proof

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

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

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

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

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

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

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

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

■

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

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

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

Proof. We simply observe that

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

■

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

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

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]

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 for μ 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

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

with $\mathbb{E}[X] = 0$ and $\operatorname{Var}[X] \leq 1$. 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 $X_1, \ldots, X_n$ 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

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

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

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 Possion 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. ⊛

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

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

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 (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 SubExp(σ^2, α) random variables have the same bound on their MGF as a Subg(σ^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

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

⊛

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

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

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

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 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 $X \in \mathrm{SubExp}(\sigma^2, \alpha)$ 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 $X_i \in \mathrm{SubExp}(\sigma_i^2, \alpha_i)$ be all independent with mean μ_i, then

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

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 $X_i \sim \mathrm{SubExp}(\sigma_i^2, \alpha_i)$ 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 $X_i \sim \mathrm{SubExp}(\sigma_i^2, \alpha_i)$ 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: Applications of Bernstein's Inequality

2.5 Bounded Difference Concentration Inequality

28 Aug. 9:00

Now we see some applications of Bernstein's inequality to bounded random variables, addressing some weaknesses of Hoeffding's inequality. Specifically, the celebrating bounded difference concentration inequality is proved.

Lemma 2.5.1. Let $|X-\mu| \leq b$ and $X-\mu$ is $\mathrm{Subg}(b^2)$. It's also true that $\mathrm{SubExp}(2\sigma^2, 2b)$.

Proof. We have

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

where we have used $(X-\mu)^k \leq (X-\mu)^2|X-\mu|^{k-2} \leq (X-\mu)^2 b^{k-2}$. 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 $1+x \leq e^x$, then the whole thing is less or equal to

$$e^{\frac{\lambda^2\sigma^2}{2(1-b|\lambda|)}} \leq e^{\lambda^2\sigma^2}$$

if $|\lambda| < 1/2b$. ■

From this, we also get $\sum_i (X_i - \mu_i) \in \mathrm{SubExp}(2\sum_i \sigma_i^2, 2b)$, i.e., from

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

we similarly have

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

Corollary 2.5.1. Let $X_1, \dots, X_n$ be independent random variables with $\mathbb{E}[X_i] = \mu_i$, $\mathrm{Var}[X_i] = \sigma^2$, and $|X_i - \mu_i| \leq b$, we have

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

Remark. If $t \leq 3\sigma^2/b$, we get back the sub-gaussian tail; if $t > 3\sigma^2/b$, we then need to look at $bt/3$ and get the sub-exponential tail.

Proposition 2.5.1.

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

Proof. Let

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

then

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

■

Example. Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Bern}(p)$, and $\hat{p} = \overline{X} = 0$. Then $p \leq \sqrt{p}/\sqrt{n}\cdots + p\dots$, i.e., $p \leq O(\frac{1}{n})$.

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

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

Theorem 2.5.1 (Bounded difference concentration inequality)**.** Let $X_1, \dots, X_n$ be i.i.d. random variables on χ, and let $g\colon \chi^n \to \mathbb{R}$ satisfying

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

for all i, then

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

The same bound holds for the left tail.

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

Remark. This is a generalization of Hoffding's inequality, where we let

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

for $X_i \in [a_i, b_i]$. In this case, we have $c_i = (b_i - a_i)/n$. Plugging in, we get back Hoffding's inequality.

Now, one can show

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

Then from Theorem 2.5.1,

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

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

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

⊛

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

Lecture 5: Proof of the Bounded Difference Concentration Inequality

We're now almost ready to prove the bounded difference concentration inequality. We first note the following. 1 Sep. 9:00

Note. The condition

$$\sup_{x_1, \ldots, x_n, x_i'} |f(x_1, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i$$

is equivalent as

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

Remark. It's useful to go through the proof of martingale decomposition and Azuse-Hoeffding inequality.

Now, we start our proof of bounded difference concentration inequality.

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

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

Now, observe that it's enough to show

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

Note. We will only show

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

however. To obtain the whole proof, we need MG decomposition.

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

Notation. Let $X^{(i)} := (X_1', \ldots, X_i', X_{i+1}, \ldots, X_n)$, and $X^{[i]} := (X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n)$.

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

$$\begin{aligned}
M'(\lambda) &= \mathbb{E}\left[f(X) - e^{\lambda f(X)}\right] \\
&= \mathbb{E}\left[f(X) - f(X')e^{\lambda f(X)}\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n} (f(X^{(i-1)}) - f(X^{(i)})) \cdot e^{\lambda f(X)}\right] \\
&= \mathbb{E}\left[\frac{1}{2}\sum_{i=1}^{n}\left(f(X^{(i-1)}) - f(X^{(i)})\right) \cdot \left(e^{\lambda f(X)} - e^{\lambda f(X^{[i]})}\right)\right].
\end{aligned}$$

Claim. For all $x \neq y \in \mathbb{R}$,

$$\frac{e^x - e^y}{x - y} \leq \frac{e^x + e^y}{2},$$

hence

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

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

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

⊛

With this, we further have

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

■

2.5.1 U-Statistics

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

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

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

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

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

Assuming 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

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

2.5.2 Beyond the Bounded Difference Concentration Inequality

Let's see some more advanced inequalities. Recall our notation.

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. Let $X = (X_1, \ldots, X_n)$, then

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

Remark.

- $\operatorname{Var}[X] = \frac{1}{2}\mathbb{E}[X - X']^2$.
- $f(X_1, \ldots, X_n) = \sum_i X_i$
- If f satisfies bounded condition, then $\operatorname{Var}[f] \leq \frac{1}{2}\sum_i c_i^2$.

Notation (Empirical process notation)**.** $\mathbb{P}f$, $\mathbb{P}_n$ be the empirical measure, then

$$
\frac{1}{n}\sum_i f(X_i) = \mathbb{P}_n f.
$$

Recall that

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

with probability at least $1-\delta$, where we assume $\mathscr{F} \ni f$ is B-bounded. If $\operatorname{Var}\left[f(X)\right]$ is very small, then we would want to replace B in the inequality by $\operatorname{Var}\left[f(X)\right]$.

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

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

Chapter 3

Lecture 6

6 Sep. 9:00

As previously seen (Uniform law of large numbers)**.** Goodness of Fit Testing: Given $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, we want to distinguish between $H_0 \colon P = P_0$ and $H_1 \colon P \neq P_0$. Many tests are possible. One approach could be that

D_n should not converge to 0, under some alternative. Assuming continuity of F, Kolmogorov showed that the distribution of D_n does not depend on F, $D_n = O_p(1/\sqrt{n})$, and $\sqrt{n} D_n \to \sup_{t\in[0,1]} \lvert B(t) \rvert$ where $B(t)$ is the Broweian Bridge on $[0,1]$. We'll take a non-asymptotic approach to this problem.

3.1 Statistical Learning and Empirical Risk Minimization

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

$$\ell \colon \mathscr{Y} \times \mathscr{Y} \to \mathbb{R}$$

and a class of functions

$$\mathscr{F} = \{ f \colon \chi \to \mathscr{Y} \},$$

the *empirical risk minimization* is defined as the problem of finding

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

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

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

Example (Linear classification)**.** Consider $\chi = \mathbb{R}^d$ and $\mathscr{Y} = \mathbb{R}$, with $\mathscr{F} = \{x \to (\operatorname{sgn}(w^\top x) + 1)/2 \colon w \in B_2^d\}$ and $\ell(a,b) = \mathbb{1}_{a \neq b}$.

Now, let's consider the expected loss (population/test error) of f be $L(f) = \mathbb{E}_{(X,Y)\sim\mathbb{P}}\left[\ell(f(x), y)\right]$, and the empirical loss

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

Problem 3.1.2. What is an upper-bound on the expected loss of ERM?

We see that while $\hat{f}$ is a function of S, $L(\hat{f}) = \mathbb{E}_{(X,Y)}\left[\ell(\hat{f}(x), Y)\right]$ is a random variable of S.

Lemma 3.1.1. For any $\mathscr{F}$, the ERM $\hat{f}$ satisfies

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

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

hence

$$L(\hat{f}) - L(f^*) \leq L(\hat{f}) - \hat{L}(\hat{f}) \leq \sup_{f\in\mathscr{F}} (L(f) - \hat{L}(f)).$$

> **Note.** We can't say $\mathbb{E}[L(\hat{f}) - \hat{L}(\hat{f})] = 0$ since $\hat{f}$ is also random.

■

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

Remark. For "curved" loss function 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 (1D classification with thresholds). Let $\ell(a, b) = \mathbb{1}_{a\neq b} = a + (1-2a)b$ for $a, b \in \{0, 1\}$. Then

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

which can be viewed essentially as[a]

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

If

$$\mathscr{F} = \{x \to \mathbb{1}_{x\leq\theta} \colon \theta \in \mathbb{R}\},$$

then

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

[a] While $Y - \sum_i y_i/n$ is independent of f so can be taken out, $1 - 2Y$ is the sign which can also be dropped essentially.

Appendix

Bibliography

[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).