STAT576
Empirical Process Theory

Pingbang Hu

August 25, 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

2 Bounding Supremum of Empirical Process **5**
2.1 Concentration Inequalities 5

Chapter 1

Introduction

Lecture 1: Introduction to Mathematical Statistics

1.1 What is Empirical Process Theory?

21 Aug. 9:00

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

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

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

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

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

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

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

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

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

Definition 1.1.2 (Empirical process)**.** Let 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

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

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]
> - $\mathscr{F} = \{\mathbb{1}_{[-\infty,t]} : t \in \mathbb{R}\}$
> - $U_n(t) = \sqrt{n}(F_n(t) - t)$ 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 $t \in [0,1]$, $U_n(t) \to \mathcal{N}(0, t - t^2)$ 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 $\{U_n(t) : t \in [0,1]\}$ 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 $X_1, \dots, X_n$ 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 $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$. In the asymptotic framework, the two key questions are the following.

Problem. Is $\hat{\theta}$ consistent for 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.

Chapter 2

Bounding Supremum of Empirical Process

Lecture 2: Concentration Inequalities and the MGF Trick

Most of this course will focus on bounding suprema of the empirical process. Let's first remind what the goal is: 23 Aug. 9:00

> **As previously seen.** Given a domain χ, probability measure $\mathbb{P}$ on χ, data $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, and a function class $\mathscr{F} \ni f \colon \chi \to \mathbb{R}$. We would like to bound (non-asymptotically)
>
> $$S_n(\mathscr{F}) = \sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}\left[f(X)\right] \right|.$$

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.

2.1 Concentration Inequalities

2.1.1 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. This is given in Lemma 2.1.1.

> **Lemma 2.1.1.** For $Z \sim \mathcal{N}(0,1)$,
>
> $$\left(\frac{1}{t} - \frac{1}{t^3}\right) \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \leq \mathbb{P}(Z \geq t) \leq \frac{1}{t} \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2}.$$

> **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. ⊛

2.1.2 MGF Trick

The most basic tool to bound tail probabilities is the Markov's inequality.

Lemma 2.1.2 (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; 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 (moment generating function) exists, like for the Gaussian, then we can expect faster tail decay. ⊛

For example, if we assume second moment exists, then we can get a $O(1/t^2)$ tail decay by Chebyshev inequality.

Lemma 2.1.3 (Generalized Chebyshev inequality)**.**

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

Remark. For $p = 2$, we have the usual form $\mathbb{P}(|X - \mu| \geq t) \leq \frac{\mathrm{Var}[X]}{t^2}$.

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.

In the same vein, we can now assume the MGF exists and apply Markov's inequality.

Lemma 2.1.4 (MGF trick (Crarmer-Chernoff method))**.**

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

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.

2.1.3 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 2.1.1 (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

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

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

[a] Also 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 2.1.5 (Equivalent conditions). Given a random variable X with $\mathbb{E}[X] = 0$, the following are equivalent for absolute constants $c_1, \ldots, c_5 > 0$.

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

(b) $\mathbb{P}(|X| \geq t) \leq 2\exp\left(-\frac{t^2}{c_2^2}\right)$.

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

(d) For all λ such that $|\lambda| \leq 1/c^4$, $\mathbb{E}\left[e^{\lambda X^2}\right] \leq e^{c_4^2\lambda^2}$.

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

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

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

where the last inequality is obtained by 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

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

since $(2k)! \geq 2^k \cdot k!$. ⊛

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

Lemma 2.1.6. Given $X \in [a,b]$ such that $\mathbb{E}[X] = 0$. Then

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

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

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

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

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

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

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

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

■

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

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

Just like independent Gaussians, sums of independent sub-gaussians remain sub-gaussian.

Lemma 2.1.7 (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

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

■

We can now immediately prove the Hoeffding's inequality.

Theorem 2.1.1 (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

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

for all $t \ge 0$.[a]

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

Proof. It's immediate from Lemma 2.1.7 and the equivalent condition (b) in Lemma 2.1.5. ■

Lecture 3

25 Aug. 9:00

Corollary 2.1.2. If $X_i \in [a,b]$ and $\sigma_i^2 = (b-a)^2/4$, then

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

Remark. If X_i are i.i.d., then

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

We also know that

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

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

In this case, we observe that for the asymptotic one, the confidence interval would be

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

while from the Hoffding's inequality, we have

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

which is much larger.

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

We see that $\mathbb{P}(X=0) = 1 - \frac{1}{k^2}$, and $\mathbb{P}(X = \pm k) = \frac{1}{2k^2}$, i.e., if $\mathrm{Var}[X] \leq 1$, we will have range $\to 2k$. Let's see some non-examples.

Example (Non-examples). Cauchy, exponential, and Possion random variables are not sub-gaussians.

Problem. What about mixture? Let's say

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

where both $Z_1, Z_2 \sim \mathrm{Subg}(\sigma)$. Is this a sub-gaussian random variable?

2.1.4 Sub-Exponential Random Variables

Let $Z^2 \sim \chi^2$, then $\mathbb{P}(Z^2 > t) = 2\mathbb{P}(Z \geq \sqrt{t}) \leq 2e^{-t/2}$. Then,

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

Definition 2.1.2 (Sub-exponential). X is *sub-exponential* with parameters (σ^2, α) with mean λ if

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

for all $|\lambda| < 1/\alpha$.

Example. For $Z^2 \sim \chi^2$, $Z^2 \sim \text{SubExp}(2, 4)$.

Proof. For all $|\lambda| < 1/4$, we have

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

By Definition 2.1.2, we're done. ⊛

Lemma 2.1.8. $X \sim \text{SubExp}(\sigma^2, \alpha)$ with mean μ. Then

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

Proof. We see that

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

for all $0 \leq \lambda < 1/\alpha$. Then, from elementary algebra, we see that

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

■

Remark. $\mathbb{P}(|X - \mu| \geq t) \leq 2\exp\left(-\min\left\{\frac{t^2}{2\sigma^2}, \frac{t}{2\alpha}\right\}\right)$. Or by observing $\min(1/u, 1/v) \geq 1/(u+v)$, we have

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

for all $t \geq 0$.

Lemma 2.1.9. If $X_i \sim \text{SubExp}(\sigma_i^2, \alpha_i)$ are all independent, then

$$\sum_{i=1}^n (X_i - \mu_i) \sim \text{SubExp}\left(\sum_i \sigma_i^2, \|\alpha\|_\infty\right).$$

Proof. Since

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

for $|\lambda| < 1/\|\alpha\|_\infty$. ■

Theorem 2.1.2 (Bernstein's inequality). Let $X_i \sim \text{SubExp}(\sigma^2, \alpha)$ are all independent, then

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

Furthermore, let $k \geq \sigma_i, \alpha_i$ for all i, then for all $a_i \in \mathbb{R}$, we have

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

Note. For $a_i = 1/\sqrt{n}$,

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

Application of Bernstein to bounded random variable.

Lemma 2.1.10. Let $|X - \mu| \leq b$ and $X - \mu$ is Subg(b^2). It's also true that SubExp$(2\sigma^2, 2b)$.

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).