High Dimensional Probability and Statistics **2nd Semester, 2021-2022**

Lecture 1: Chernoff Method and Concentration Inequalities

Instructor: Ke Wei *Scribe: Ke Wei (Updated: 2022/02/26)*

Agenda:

- Preliminaries
- Sub-Gaussian distributions and Hoeffding inequality
- Sub-exponential distributions and Bernstein inequality
- Martingale methods and bounded differences inequality

1.1 Preliminaries

Theorem 1.1 *Let X be a non-negative random variable. Then,*

$$\mathbb{E}\left[X\right] = \int_0^\infty \mathbb{P}\left[X > t\right] dt.$$

Proof: We have

$$\mathbb{E}\left[X\right] = \mathbb{E}\left[\int_0^\infty 1_{\{t<X\}} dt\right] = \int_0^\infty \mathbb{E}\left[1_{\{t<X\}}\right] dt = \int_0^\infty \mathbb{P}\left[X > t\right] dt,$$

as claimed. ■

Exercise 1.2 *Let X be a random variable and $p \in (0, \infty)$. Show that*

$$\mathbb{E}\left[|X|^p\right] = \int_0^\infty p t^{p-1} \mathbb{P}\left[|X| > t\right] dt.$$

Theorem 1.3 (Jensen's inequality) *If f is convex, then*

$$\mathbb{E}\left[f(X)\right] \geq f(\mathbb{E}\left[X\right]).$$

If f is concave, then

$$\mathbb{E}\left[f(X)\right] \leq f(\mathbb{E}\left[X\right]).$$

Proof: It suffices to prove the first inequality. Let $l(x)$ be the tangent line of $f(x)$ at $\mathbb{E}\left[X\right]$. Then,

$$\mathbb{E}\left[f(X)\right] \geq \mathbb{E}\left[l(X)\right] = l(\mathbb{E}\left[X\right]) = f(\mathbb{E}\left[X\right]),$$

where the first equality follows from the fact that $l(X)$ is a linear function and the second equality follows from that $l(x)$ is tangent to $f(x)$ at $\mathbb{E}\left[X\right]$. ■

Example 1.4 *Let $X_1, \cdots, X_n$ be standard Gaussian random variables, i.e., $X_k \sim \mathcal{N}(0, \sigma^2)$. Then,*

$$\mathbb{E}\left[\max_{k=1,\cdots,n} X_k\right] \leq \sigma\sqrt{2\log n}.$$

Proof: First note that for any $\lambda > 0$

$$\begin{aligned}
\mathbb{E}\left[\exp(\lambda X_k)\right] &= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}\exp(\lambda x)\exp(-x^2/2\sigma^2)dx \\
&= \exp(\sigma^2\lambda^2/2)\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}\left(\frac{x}{\sigma}-\sigma\lambda\right)^2\right)dx \\
&= \exp(\sigma^2\lambda^2/2). \qquad (1.1)
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda\max_k X_k\right)\right] = \mathbb{E}\left[\max_k \exp\left(\lambda X_k\right)\right] &\leq \sum_k \mathbb{E}\left[\exp\left(\lambda X_k\right)\right] \\
&\leq n\exp(\sigma^2\lambda^2/2) = \exp(\log(n) + \sigma^2\lambda^2/2).
\end{aligned}$$

Thus, the application of Jensen's inequality yields that

$$\exp\left(\mathbb{E}\left[\lambda\max_k X_k\right]\right) \leq \mathbb{E}\left[\exp\left(\lambda\max_k X_k\right)\right] \leq \exp(\log(n) + \sigma^2\lambda^2/2),$$

which leads to

$$\mathbb{E}\left[\max_k X_k\right] \leq \frac{\log(n)}{\lambda} + \frac{\sigma^2\lambda}{2}.$$

Taking $\lambda = \sqrt{2\log(n)}/\sigma$ concludes the proof. ■

One goal of this course is to study the concentration of a random quantity (function of random variables/vectors/matrices) around its mean.

There are two probability results which can provide some instructions: law of large numbers and central limit theorem. As an example, consider the problem of estimating the unknown mean μ of a random variable from its i.i.d samples $X_1, \cdots, X_n$. A very natural way to estimate μ is to use the sample mean

$$f(X_1, \cdots, X_n) = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

It is not hard to see that $\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n} X_k\right] = \mu$. Moreover, the law of large numbers tells us that

$\frac{1}{n}\sum_{k=1}^{n} X_k$ *converges to μ in probability when n goes to infinity.*

If the variance of the random variable exists, denoted σ^2, the central limit theorem implies that

$$\frac{\sum_{k=1}^n (X_k - \mu)}{\sigma\sqrt{n}} \sim \mathcal{N}(0,1) \textit{ when } n \textit{ goes to infinity.}$$

However, both the law of large number and the central limit theorem are asymptotic results and they cannot tell us how well $f(X_1, \cdots, X_n) = \frac{1}{n}\sum_{k=1}^n X_k$ concentrates around μ when n is finite. That is, they cannot be used to bound the probability that $\frac{1}{n}\sum_{k=1}^n X_k$ deviates from its mean, namely

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{k=1}^n X_k - \mu\right| \geq t\right]. \tag{1.2}$$

In this course, we will study this kind of concentration problem directly. Indeed, we will begin with the bound for (1.2), and then move to concentration inequalities which involve more complicated f than the sum of random variables in order to tackle more difficult problems than mean estimation.

1.1.1 Markov and Chebshev Inequalities

One way to bound the tail probability of a random variable is to control its moments. The most elementary tail bound is Markov inequality which only uses the expectation (first moment) of the variable.

Theorem 1.5 (Markov inequality) *If X is a non-negative variable, then any $t > 0$ one has*

$$\mathbb{P}[X > t] \leq \frac{\mathbb{E}[X]}{t}.$$

Proof: A simple calculation yields that

$$\mathbb{E}[X] \geq \mathbb{E}\left[X 1_{\{X>t\}}\right] \geq t\mathbb{P}[X > t],$$

as claimed. ■

Using high order moments typically leads to stronger probability bounds, e.g., Chebshev inequality.

Theorem 1.6 (Chebshev inequality) *For a random variable with finite variance, there holds,*

$$\mathbb{P}[|X - \mathbb{E}[X]| > t] \leq \frac{\mathbb{E}\left[|X - \mathbb{E}[X]|^2\right]}{t^2}.$$

Proof: Apply Markov inequality to the random variable $|X - \mathbb{E}[X]|^2$ ■

Example 1.7 *Let X be a Bernoulli variable,*

$$X = \begin{cases} 1 & \textit{with probability } p \\ 0 & \textit{with probability } 1-p. \end{cases}$$

Let $X_k,\ i=1,\cdots,n$ be i.i.d copies of X, and define $S_n=\sum_{k=1}^{n}X_k$. For a positive number $p<\alpha<1$, the application of Markov inequality gives

$$\mathbb{P}\left[S_n>\alpha n\right]\le\frac{\mathbb{E}\left[S_n\right]}{\alpha n}=\frac{p}{\alpha},$$

while the application of Chebshev inequality gives

$$\begin{aligned}\mathbb{P}\left[S_n>\alpha n\right]&=\mathbb{P}\left[S_n-pn>(\alpha-p)n\right]\\&\le\mathbb{P}\left[|S_n-pn|>(\alpha-p)n\right]\\&\le\frac{\mathbb{E}\left[|S_n-pn|^2\right]}{(\alpha-p)^2n^2}\\&=\frac{p(1-p)}{(\alpha-p)^2n}.\end{aligned}$$

This example shows that we can have a better bound (order of $1/n$ rather than a constant order) by Chebshev inequality. As can be seen later, by one of the main results in this lecture – Hoeffding inequality, we can establish a tail bound that decays exponentially fast.

There is a natural way to extend the Markov inequality to random variables with higher-order moments. For instance, if $\mathbb{E}\left[|X-\mathbb{E}\left[X\right]|^k\right]$ exists for some $k>1$, then an application of the Markov inequality to the random variable $|X-\mathbb{E}\left[X\right]|^k$ yields that

$$\mathbb{P}\left[|X-\mathbb{E}\left[X\right]|>t\right]\le\frac{\mathbb{E}\left[|X-\mathbb{E}\left[X\right]|^k\right]}{t^k}.$$

Of course, we can use other functions rather than a single moment of the random variable. The tight bounds that will be established next are indeed based on the *moment generating function* (MGF, a mixture of all moments),

$$\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right].$$

In the same spirit of the Markov or Chebshev inequality, we have

$$\mathbb{P}\left[X-\mathbb{E}\left[X\right]>t\right]=\mathbb{P}\left[e^{\lambda(X-\mathbb{E}[X])}>e^{\lambda t}\right]\le e^{-\lambda t}\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right],\quad\lambda>0.$$

Note that in the above inequality, there is a free parameter $\lambda>0$ to choose. The *Chernoff method* chooses λ in an interval $[0,b]$ (b can be infinite or finite up to the bound of moment generating function) such that the righthand side is minimized, leading to

$$\mathbb{P}\left[X-\mathbb{E}\left[X\right]>t\right]\le\inf_{\lambda\in[0,b]}e^{-\lambda t}\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right].\tag{1.3}$$

It is easy to see that the key in the application of the Chernoff method is to estimate $\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right]$. Indeed, one advantage of using moment generating function over the all possible polynomials is that the former one is a smooth function with the parameter λ and can be easily manipulated. Next we will study two different distributions based on the different behaviors of their moment generating functions, as well as the corresponding concentration inequalities.

1.2 Sub-Gaussian Distributions and Hoeffding Inequality

Let $X \sim \mathcal{N}(\mu, \sigma^2)$ be a normal/Gaussian distribution of mean μ and variance σ^2. We have

$$\mathbb{P}\left[|X-\mu| \geq t\right] \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right) \tag{1.4}$$

Exercise 1.8 *Prove* (1.4)*.*

The above inequality shows the tail bound of normal distribution decays exponentially fast. Thus, it is interesting to see whether there are other distributions which exhibit similar behavior. The answer is affirmative, and this family of distributions are known as sub-Gaussian distributions. They are fully characterized by the behavior of their moment generating functions.

Definition 1.9 (Sub-Gaussian distribution) *A random variable X with mean μ is sub-Gaussian if there exists a positive number $\nu > 0$ such that*

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\lambda^2\nu^2/2} \quad \textit{for all} \quad \lambda \in \mathbb{R}. \tag{1.5}$$

Remark 1.10 *Though here ν is NOT equivalent to the variance of a random variable, we can sometimes think of it as the variance to get some intuition.*

Example 1.11 (Gaussian distribution) *Let $X \sim \mathcal{N}(\mu, \sigma^2)$ be a Gaussian random variable. It follows from* (1.1) *that*

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] = e^{\lambda^2\sigma^2/2} \;\; \textit{for all} \;\; \lambda \in \mathbb{R}.$$

Thus X is sub-Gaussian with parameter $\nu = \sigma$.

Example 1.12 (Rademacher variables) *A Rademacher random variable ε takes the values $\{-1, +1\}$ in the same probability. By taking expectations and using the power series expansion, we have*

$$\begin{aligned}\mathbb{E}\left[e^{\lambda X}\right] &= \frac{1}{2}\left(e^{-\lambda} + e^{\lambda}\right) \\ &= \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{(2k)!} \\ &\leq 1 + \sum_{k=1}^{\infty} \frac{\lambda^{2k}}{2^k k!} \\ &= e^{\lambda^2/2},\end{aligned}$$

which shows that ε is a sub-Gaussian variable with parameter $\nu = \sigma = 1$.

Example 1.13 (Bounded random variables) *Let X be zero-mean, and supported on a closed interval $[a, b]$. We claim that*

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

In other words, X is sub-Gaussian with parameter $(b-a)/2$. To show this, define $\psi(\lambda)$ (knowns as log-moment generating function) as

$$\psi(\lambda) = \log \mathbb{E}\left[e^{\lambda X}\right].$$

Then it suffices to show

$$\psi(\lambda) \leq \frac{\lambda^2(b-a)^2}{8}.$$

First, it is not hard to see that

$$\psi'(\lambda) = \frac{\mathbb{E}\left[Xe^{\lambda X}\right]}{\mathbb{E}\left[e^{\lambda X}\right]}$$

and

$$\begin{aligned}\psi''(\lambda) &= \frac{\mathbb{E}\left[e^{\lambda X}\right]\mathbb{E}\left[X^2e^{\lambda X}\right] - (\mathbb{E}\left[Xe^{\lambda X}\right])^2}{(\mathbb{E}\left[e^{\lambda X}\right])^2} \\ &= \mathbb{E}\left[X^2\frac{e^{\lambda X}}{\mathbb{E}\left[e^{\lambda X}\right]}\right] - \left(\mathbb{E}\left[X\frac{e^{\lambda X}}{\mathbb{E}\left[e^{\lambda X}\right]}\right]\right)^2.\end{aligned}$$

It follows immediately that

$$\psi'(0) = 0.$$

Moreover, the expression for $\psi''(\lambda)$ implies that $\psi''(\lambda)$ is indeed the variance of X after a change of measure. Thus, by the variational definition o variance, we have

$$\psi''(\lambda) \leq \mathbb{E}\left[\left(X - \frac{b+a}{2}\right)^2 \frac{e^{\lambda X}}{\mathbb{E}\left[e^{\lambda X}\right]}\right] \leq \frac{(b-a)^2}{4}, \quad \forall \lambda \in \mathbb{R}.$$

Also noting that $\psi(0) = 0$, we finally have

$$\psi(\lambda) = \psi(0) + \psi'(0)\lambda + \frac{1}{2}\psi''(\xi)\lambda^2 \leq \frac{\lambda^2(b-a)^2}{8},$$

which completes the proof.

1.2.1 Hoeffding Inequality

By the Chernoff method (see (1.3)) we can show that sub-Gaussian random variables have the same concentration properties as Gaussian random variables.

Theorem 1.14 (Hoeffding inequality) *Let X (with $\mathbb{E}[X] = \mu$) be a sub-Gaussian random variable with parameter ν. Then,*

$$\mathbb{P}\left[|X-\mu| > t\right] \leq 2e^{-\frac{t^2}{2\nu^2}}.$$

Proof: Inserting the sub-Gaussian property into(1.3) and optimizing the right hand side of the above inequality with respect to $\lambda > 0$ yields that

$$\mathbb{P}\left[X-\mu>t\right] \leq e^{-\frac{t^2}{2\nu^2}}.$$

Moreover, by considering $-X$, we can get

$$\mathbb{P}\left[X-\mu<-t\right] \leq e^{-\frac{t^2}{2\nu^2}},$$

which concludes the proof. ■

Chernoff bounds can be easily extended to sums of independent random variables because of the tensorization property of the moment generating functions in this situation, i.e., moment generating functions of sums of independent random variables become products of moment generating functions.

Proposition 1.15 *Let $X_1, \cdots, X_n$ be independent ν_k^2 sub-Gaussian random variables. Then $\sum_{k=1}^n X_k$ is a sub-Gaussian random variable with parameter $\nu = \sum_{k=1}^n \nu_k^2$.*

Proof: The moment generating function $\sum_{k=1}^n X_k$ can be upper bounded as

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda\left(\sum_{k=1}^n X_k - \mathbb{E}\left[\sum_{k=1}^n X_k\right]\right)\right)\right] &= \mathbb{E}\left[\prod_{k=1}^n \exp\left(X_k - \mathbb{E}\left[X_k\right]\right)\right] = \prod_{k=1}^n \mathbb{E}\left[\exp\left(X_k - \mathbb{E}\left[X_k\right]\right)\right] \\ &\leq \prod_{k=1}^n \exp\left(\frac{\lambda^2\nu_k^2}{2}\right) = \exp\left(\frac{\lambda^2\sum_{k=1}^n \nu_k^2}{2}\right),\end{aligned}$$

which completes the proof. ■

The follow general Hoeffding inequality follows immediately from Theorem 1.14 and Proposition 1.15.

Theorem 1.16 (General Hoeffding inequality) *Suppose $X_k,\ k = 1, \cdots, n$ are independent random variables, and X_k has mean μ_k and sub-Gaussian parameter ν_k. Then for all $t \geq 0$, we have*

$$\mathbb{P}\left[\left|\sum_{k=1}^n (X_k-\mu_k)\right| > t\right] \leq 2\exp\left(-\frac{t^2}{2\sum_{k=1}^n \nu_k^2}\right).$$

Example 1.17 *Suppose $X_k,\ k = 1, \cdots, n$ are independent random variables satisfying $\mathbb{E}\left[X_k\right] = \mu_k$ and $a \leq X_k \leq b$. Then for all $t \geq 0$, we have*

$$\mathbb{P}\left[\left|\sum_{k=1}^n (X_k-\mu_k)\right| > t\right] \leq 2\exp\left(-\frac{2t^2}{n(b-a)^2}\right).$$

Example 1.18 *Let us revisit Example 1.7 using the Hoeffding inequality, yielding*

$$\mathbb{P}\left[S_n > \alpha n\right] = \mathbb{P}\left[\sum_{k=1}^n (X_k - p) \geq (\alpha-p)n\right] \leq \exp\left(-\frac{(\alpha-p)^2 n}{2}\right),$$

which decreases faster than what Chebshev inequality gives.

1.2.2 Equivalent Characterizations of sub-Gaussian Distribution[1]

We have shown that the sub-Gaussian property implies the exponential decay of the tail probability. In fact, the converse direction also holds true. Moreover, there are several equivalent characterizations of the sub-Gaussian distribution.

Theorem 1.19 *Let X be a mean zero random variable. Then the following four statements are equivalent.*

1. *X is sub-Gaussian satisfying,*

$$\mathbb{E}\left[\exp\left(\lambda X\right)\right] \leq \exp\left(c_1\lambda^2\nu^2\right) \quad \textit{for all} \quad \lambda \in \mathbb{R}.$$

2. *The tails of X satisfy*

$$\mathbb{P}\left[|X| \geq t\right] \leq 2\text{exp}\left(-\frac{t^2}{c_2\nu^2}\right) \quad \textit{for all } t \geq 0.$$

3. *The moments of X satisfy*

$$\|X\|_{L_p} := \left(\mathbb{E}\left[|X|^p\right]\right)^{1/p} \leq c_3\nu\sqrt{p} \quad \textit{for all} \quad p \geq 1.$$

4. *The moment generating function of X^2 is bounded at some point[2],*

$$\mathbb{E}\left[\exp\left(\frac{X^2}{c_4\nu^2}\right)\right] \leq e.$$

Here, $c_i, \ i = 1, \cdots, 4$ are positive, absolute constants (see the notational remark in the syllabus).

Proof: We will proceed the proof in the following way: $1 \Rightarrow 2 \Rightarrow 3 \Rightarrow 4 \Rightarrow 1$.

$1 \Rightarrow 2$: We have established this above using the Chernoff method.

$2 \Rightarrow 3$: W.l.og, assume $c_2 = 1$. Then,

$$\begin{aligned}
\mathbb{E}\left[|X|^p\right] &= p\int_0^\infty t^{p-1}\mathbb{P}\left[|X| \geq t\right]dt \\
&\leq 2p\int_0^\infty t^{p-1}\text{exp}\left(-\frac{t^2}{\nu^2}\right)dt \\
&= p\nu^p\int_0^\infty s^{\frac{p}{2}-1}e^{-s}ds \qquad (\text{letting } s = \frac{t^2}{\nu^2}) \\
&= p\nu^p\Gamma(p/2) \qquad (\Gamma(z) \text{ is a Gamma function}) \\
&\leq p\nu^p\left(p/2\right)^{p/2} \qquad (\Gamma(z) \leq z^z, \textbf{ check this! }).
\end{aligned}$$

Taking the p-th root on both sides and noting that $p^{1/p} \leq e$ (**check this!**) concludes the proof.

[1] This part can be skipped if you find it difficult.

[2] The constant e on the righthand side does not have any special meaning and can be replaced by any absolute constant (similar to different scales of a norm.

3 ⇒ 4: As above, we can assume $c_3 = 1$. Then

$$\begin{aligned}\mathbb{E}\left[\exp\left(\frac{X^2}{c_4\nu^2}\right)\right] &= \sum_{p=0}^{\infty}\frac{\mathbb{E}\left[X^{2p}\right]}{p!c_4^p\nu^{2p}} \le \sum_{p=0}^{\infty}\frac{\nu^{2p}(2p)^p}{p!c_4^p\nu^{2p}}\\ &\le \sum_{p=0}^{\infty}\left(\frac{2e}{c_4}\right)^p = \frac{1}{1-2e/c_4} \le e \qquad (\text{use } p! \ge (p/e)^p, \textbf{ check this!})\end{aligned}$$

provided $c_4 \ge 2e/(1-1/e)$.

4 ⇒ 1: Again, we can assume $c_4 = 1$. First noting that

$$\lambda x \le \frac{\lambda^2\nu^2}{2} + \frac{x^2}{2\nu^2},$$

we have

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda X\right)\right] &\le \exp\left(\lambda^2\nu^2/2\right)\mathbb{E}\left[\exp\left(X^2/(2\nu^2)\right)\right] \le \exp\left(\lambda^2\nu^2/2\right)\sqrt{\mathbb{E}\left[\exp\left(X^2/(\nu^2)\right)\right]}\\ &\le e^{1/2}\exp\left(\lambda^2\nu^2/2\right) \le \exp\left(\lambda^2\nu^2\right)\end{aligned}$$

provided $|\lambda| \ge 1/\nu$, where the second inequality follows from the Jensen inequality to the function $\sqrt{x}$. Thus, it remains to discuss the case $|\lambda| < 1/\nu$. In this situation, using the inequality $e^x \le x + e^{x^2}$ (**check this!**) we have

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda X\right)\right] &\le \underbrace{\mathbb{E}\left[\lambda X\right]}_{=0} + \mathbb{E}\left[\exp\left(\lambda^2X^2\right)\right] = \mathbb{E}\left[\exp\left(\lambda^2X^2\right)\right] = \mathbb{E}\left[\left(\exp\left(X^2/\nu^2\right)\right)^{(\lambda^2\nu^2)}\right]\\ &\le \left(\mathbb{E}\left[\exp\left(X^2/\nu^2\right)\right]\right)^{\lambda^2\nu^2}\\ &\le \exp\left(\lambda^2\nu^2\right),\end{aligned}$$

where in the second inequality we utilize the Jensen inequality by noting that $\lambda^2\nu^2 < 1$. ■

Exercise 1.20 (Khintchine inequality) *Let X_k, $k = 1, \cdots, n$ be i.i.d, zero mean, unit variance sub-Gaussian random variables with parameter ν^2. Letting $a = (a_1, \cdots, a_n) \in \mathbb{R}^n$, show that for any $p \in [2, \infty)$ we have*

$$\|a\|_2 \le \|\sum_{k=1}^{n} a_kX_k\|_{L_p} \lesssim \nu\sqrt{p}\|a\|_2.$$

(See the notational remark in the syllabus for the meaning of $\lesssim$.)

At the end of this section we present the following lemma, where a very useful *decoupling technique* via the introduction of an independent random variable for auxiliary randomness is used in the proof. See Chapter 6.1 of of *High-dimensional probability: An introduction with applications in data science* by Roman Vershynin for the general decoupling technique.

Lemma 1.21 *Let X be mean zero sub-Gaussian random variable with parameter ν^2. Then*

$$\mathbb{E}\left[\exp\left(\lambda X^2\right)\right] \le \frac{1}{[1-2\lambda\nu^2]_+^{1/2}},$$

where the equality holds for $X \sim \mathcal{N}(0, \nu^2)$.

Proof: When $X \sim \mathcal{N}(0, \nu^2)$, we can establish the equality by direction integral based on the pdf of the Gaussian distribution.

For a general sub-Gaussian variable X, let Z be an independent $\mathcal{N}(0,1)$ random variable. Noting that

$$\mathbb{E}\left[\exp\left(\lambda x Z\right)\right] = \exp\left(\frac{\lambda^2 x^2}{2}\right),$$

we have

$$\mathbb{E}\left[\exp\left(\lambda X^2\right)\right] = \mathbb{E}\left[\exp\left(\sqrt{2\lambda} X Z\right)\right] \leq \mathbb{E}\left[\exp\left(\lambda \nu^2 Z^2\right)\right] \leq \frac{1}{[1-2\lambda\nu^2]_+^{1/2}},$$

where the first inequality follows from the sub-Gaussian property of X and the second inequality follows from the the fact Z is $\mathcal{N}(0,1)$. ■

1.3 Sub-exponential Distributions and Bernstein Inequality

As we have seen from above, sub-Gaussian distribution is an extension of the Gaussian distribution. In contrast, sub-exponential distribution is an extension of the squared Gaussian distribution. For simplicity, let $X \sim \mathcal{N}(0,1)$ be standard normal distribution and let $Z = X^2$ be χ^2. Then,

$$\mathbb{E}\left[e^{\lambda(Z-1)}\right] = \begin{cases} \frac{e^{-\lambda}}{\sqrt{1-2\lambda}}, & \text{if } \lambda < \frac{1}{2} \\ \text{not exist}, & \text{otherwise.} \end{cases}$$

Thus, the moment generating function does not exist over the entire real line. Moreover, since $1 - x > e^{-x^2-x}$ (**check this!**) for all $x < 1/2$, one has

$$\mathbb{E}\left[e^{\lambda(Z-1)}\right] \leq e^{4\lambda^2/2} \text{ for all } |\lambda| < \frac{1}{4}.$$

Compared with (1.5), we see that similar bound only holds in a local neighborhood of zero. This kind of condition defines the family of sub-exponential distributions.

Definition 1.22 (Sub-exponential distribution) *A random variable X with mean μ is sub-exponential if there are non-negative parameters (ν, b) such that*

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\nu^2\lambda^2/2} \quad \textit{for all} \quad |\lambda| < 1/b.$$

Example 1.23 (χ^2-distribution) *We have shown that if $X \sim \mathcal{N}(0,1)$, then X^2 is sub-exponential with parameters $(\nu, b) = (2, 4)$.*

Example 1.24 (Exponential distribution) *Recall that X has exponential distribution with rate $a > 0$ if the pdf of X is given by*

$$f(x) = \begin{cases} ae^{-ax} & x \geq 0 \\ 0 & x < 0. \end{cases}$$

A direct calculation shows that $\mathbb{E}[X] = \frac{1}{a}$*. For simplicity let* $a = 1$*. Then we have*

$$\mathbb{E}[\exp(\lambda(X-1))] = \int_0^\infty e^{\lambda(x-1)} e^{-x} dx = \begin{cases} \frac{e^{-\lambda}}{1-\lambda} & \lambda < 1 \\ \infty & \lambda \geq 1. \end{cases}$$

The application of $1 - x > e^{-x^2-x}$ *for* $x < 1/2$ *yields that*

$$\mathbb{E}[\exp(\lambda(X-1))] \leq e^{\lambda^2} \quad \textit{for all} \quad |\lambda| < \frac{1}{2}.$$

Bernstein condition based on the moments of X provides an indirect way to verify the sub-exponential property. More precisely, let X be random variable with mean μ and variance σ^2. We say Bernstein's condition with parameter b holds if

$$\left|\mathbb{E}\left[(X-\mu)^k\right]\right| \leq \frac{1}{2} k! \sigma^2 b^{k-2} \quad \text{for } k = 3, 4, \cdots$$

Lemma 1.25 *If* X *satisfies the Bernstein condition, then* X *is sub-exponential with parameters* $(\sqrt{2}\sigma, 2b)$*.*

Proof: We have

$$\begin{aligned} \mathbb{E}\left[e^{\lambda(X-\mu)}\right] &= \sum_{k=0}^{\infty} \frac{\mathbb{E}\left[\lambda^k (X-\mu)^k\right]}{k!} \\ &\leq 1 + \frac{\sigma^2\lambda^2}{2} + \frac{\sigma^2\lambda^2}{2} \sum_{k=1}^{\infty} (|\lambda| b)^k \\ &= 1 + \frac{\sigma^2\lambda^2}{2} + \frac{\sigma^2\lambda^2|\lambda|b}{2(1-|\lambda|b)} \qquad \left(\forall\ |\lambda| < \frac{1}{b}\right) \\ &= 1 + \frac{\sigma^2\lambda^2/2}{1-|\lambda|b} \\ &\leq e^{\frac{\sigma^2\lambda^2/2}{1-|\lambda|b}} \qquad (1.6) \\ &\leq e^{\frac{\sigma^2(\sqrt{2}\lambda)^2}{2}} \quad \forall\ |\lambda| \leq \frac{1}{2b}, \end{aligned}$$

which implies X is sub-exponential with parameters $(\sqrt{2}\sigma, 2b)$. ■

Exercise 1.26 *Let* X *be a random variable with* $\mathbb{E}[X] = \mu$*. Suppose* $|X - \mu| \leq b$*. Show that* X *satisfies the Bernstein condition.*

1.3.1 Bernstein Inequality

For sub-exponential distributions we can establish the Bernstein tail, which mixes the Gaussian tail and the exponential tail.

Theorem 1.27 (Bernstein inequality) *Suppose* X *is a sub-exponential variable with parameters* (ν, b)*. Then*

$$\mathbb{P}[|X-\mu| > t] \leq 2\exp\left(-\frac{1}{2}\min\left(\frac{t^2}{\nu^2}, \frac{t}{b}\right)\right) = \begin{cases} 2e^{-\frac{t^2}{2\nu^2}}, & \textit{if } \ 0 \leq t \leq \frac{\nu^2}{b} \\ 2e^{-\frac{t}{2b}} & \textit{if } \ t > \frac{\nu^2}{b}. \end{cases}$$

Proof: We assume without loss of generality $\mu = 0$. The application of the Chernoff approach yields that

$$\mathbb{P}[X - \mu > t] \leq e^{-\lambda t}\mathbb{E}\left[e^{\lambda X}\right] \leq e^{-\lambda t + \nu^2\lambda^2/2}, \quad \forall 0 < \lambda \leq 1/b.$$

Optimizing the right hand side with respect to λ over $(0, 1/b]$ gives the one-sided tail bound. Consider $-X$ for the other tail bound. ■

Example 1.28 *Let X be a random variable such that $|X - \mu| \leq b$. We know that it is also sub-exponential with parameters $(\sqrt{2}\sigma, b)$ where σ is the variance of X. Then the Bernstein inequality implies that*

$$\mathbb{P}[|X - \mu| > t] \leq \begin{cases} 2e^{-\frac{t^2}{4\sigma^2}}, & \text{if } \; 0 \leq t \leq \frac{\sigma^2}{b} \\ 2e^{-\frac{t}{2b}} & \text{if } \; t > \frac{\sigma^2}{b}, \end{cases}$$

while the application of the Hoeffding type bound gives

$$\mathbb{P}[|X - \mu| > t] \leq 2e^{-\frac{t^2}{2b^2}}.$$

It is evident that when t is sufficiently large, the Hoeffding type bound is better than the Bernstein type bound. However, it is worth noting that if t is small, the Bernstein type bound might be better than the Hoeffding type bound since it is possible that $\sigma^2 \ll b^2$.

For sub-exponential variable satisfying the Bernstein condition, we can actually establish the following slightly improved bound

$$\mathbb{P}[|X - \mu| > t] \leq 2\text{exp}\left(-\frac{t^2}{2(\sigma^2 + bt)}\right). \tag{1.7}$$

Exercise 1.29 *Prove*(1.7)*. (**Hint:** Apply the Chernoff method to the inequality* (1.6) *directly.)*

The Bernstein inequality shows that for small t the tail bound is of the Gaussian type while for large t the tail bound is of the exponential type. Though the region of the Gaussian tail is a bit restrictive for a single random variable (for example when t is very close to 0 we even have $e^{-t} \lesssim e^{-t^2}$), for the sum of independent random variables this region will increase as the number of random variables increases.

Proposition 1.30 *Suppose that $X_k, \; k = 1, \cdots, n$ are n independent variables, and that X_k is sub-exponential with parameters (ν_k, b_k). Then $\sum_{k=1}^n (X_k - \mu_k)$ is sub-exponential with parameters (ν_*, b_*), where*

$$\nu_*^2 = \sum_{k=1}^n \nu_k^2 \quad \text{and} \quad b_* = \max_{1 \leq k \leq n} b_k.$$

Moreover, if $X_k, \; k = 1, \cdots, n$ are i.d.d sub-exponential with parameters (ν, b), then $\sum_{k=1}^n (X_k - \mu)$ is sub-exponential with parameters $(\sqrt{n}\nu, b)$.

Proof: The moment generating function of $\sum_{k=1}^n (X_k - \mu_k)$ can be bounded as follows

$$\mathbb{E}\left[\exp\left(\lambda \sum_{k=1}^n (X_k - \mu_k)\right)\right] = \prod_{k=1}^n \mathbb{E}\left[\exp\left(\lambda(X_k - \mu_k)\right)\right] \le \prod_{k=1}^n \exp\left(\lambda^2 \nu_k^2/2\right),$$

where the inequality is valid for all $\lambda < (\max_k b_k)^{-1}$. ■

The following general Bernstein inequality follows immediately from the last proposition.

Theorem 1.31 (General Bernstein inequality) *Suppose that $X_k,\ i = 1, \cdots, n$ are n independent variables, and that X_k is sub-exponential with parameters (ν_k, b_k). Then,*

$$\mathbb{P}\left[\left|\sum_{k=1}^n (X_k - \mu_k)\right| \ge t\right] \le 2\exp\left(-\frac{1}{2}\min\left(\frac{t^2}{\nu_*^2}, \frac{t}{b_*}\right)\right) = \begin{cases} 2e^{-\frac{t^2}{2\nu_*^2}}, & \text{if } 0 \le t \le \frac{\nu_*^2}{b_*} \\ 2e^{-\frac{t}{2b_*}} & \text{if } t > \frac{\nu_*^2}{b_*}, \end{cases}$$

where

$$\nu_*^2 = \sum_{k=1}^n \nu_k^2 \quad \text{and} \quad b_* = \max_{1\le k\le n} b_k.$$

Example 1.32 *Let $Z_k,\ k = 1, \cdots, n$ be i.i.d Chi-square variables. Noting that Z_k is sub-exponential with parameters $(2, 4)$, there holds*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{k=1}^n (Z_k - 1)\right| \ge t\right] \le 2\exp\left(-\frac{n}{8}\min\left(t^2, t\right)\right).$$

Additionally, as a remark we roughly illustrates that why the exponential tail in the Bernstein inequality does not contradicts the central limit theorem. Let $X_k,\ k = 1, \cdots, n$ be i.i.d sub-exponential random variables with parameters (ν, b). Consider the rescaled random variable $\frac{1}{\sqrt{n}}\sum_{k=1}^n (X_k - \mu)$. The application of the Bernstein inequality yields that

$$\mathbb{P}\left[\left|\frac{1}{\sqrt{n}}\sum_{k=1}^n (X_k - \mu)\right| \ge t\right] \le \begin{cases} 2e^{-\frac{t^2}{2\nu^2}} & 0 \le t \le \frac{\sqrt{n}\nu^2}{b} \\ 2e^{-\frac{\sqrt{n}t}{2b}} & t > \frac{\sqrt{n}\nu^2}{b}. \end{cases}$$

It is clear that the Gaussian tail bound region $0 \le t \le \frac{\sqrt{n}\nu^2}{b}$ increases linearly with respect to $\sqrt{n}$.

1.3.2 Equivalent Characterizations of sub-Exponential Distribution[3]

Under a generalized definition of sub-exponential distributions (in for example *High-dimensional probability: An introduction with applications in data science* by Roman Vershynin), we may establish the following equivalence.

Theorem 1.33 *Let X be a mean zero random variable. Then the following four statements are equivalent.*

[3] This part can be skipped if you find it difficult.

1. *X is sub-exponential satisfying,*

$$\mathbb{E}\left[\exp\left(\lambda X\right)\right] \leq \exp\left(c_1 \lambda^2 \nu^2\right) \quad \textit{for all} \quad |\lambda| \leq \frac{c_1'}{\nu}. \tag{1.8}$$

   Note that if X satisfies Definition 1.22, then it will satisfy (1.8) *with* $\max(\nu, b)$*. However, the resulting Bernstein inequality will be weaker since both ν and b will be replaced by* $\max(\nu, b)$*.*

2. *The tails of X satisfy*

$$\mathbb{P}\left[|X| \geq t\right] \leq 2\exp\left(-\frac{t}{c_2 \nu}\right) \quad \textit{for all } t \geq 0.$$

3. *The moments of X satisfy*

$$\|X\|_{L_p} = \left(\mathbb{E}\left[|X|^p\right]\right)^{1/p} \leq c_3 \nu p \quad \textit{for all} \quad p \geq 1.$$

4. *The moment generating function of $|X|$ is bounded at some point*[4]*,*

$$\mathbb{E}\left[\exp\left(\frac{|X|}{c_4 \nu}\right)\right] \leq e.$$

Here, c_i, $i = 1, \cdots, 4$ and c_1' are positive, absolute constants.

Proof: We will proceed the proof in the following way: $2 \Rightarrow 3 \Rightarrow 4 \Rightarrow 2$ and $1 \Leftrightarrow 3$.

$2 \Rightarrow 3$: W.l.o.g, we assume $c_2 = 1$. Then,

$$\begin{aligned}
\mathbb{E}\left[|X|^p\right] &= p \int_0^\infty t^{p-1} \mathbb{P}\left[|X| \geq t\right] dt \\
&\leq 2p \int_0^\infty t^{p-1} \exp\left(-t/v\right) dt \\
&= 2p \nu^p \Gamma(p) \\
&\leq 2p \nu^p p^p.
\end{aligned}$$

Taking a p-th root on both sides yields the result.

$3 \Rightarrow 4$: As above we assume $c_3 = 1$. Then,

$$\mathbb{E}\left[\exp\left(\frac{X}{c_4 \nu}\right)\right] = \sum_{p=0}^{\infty} \frac{\mathbb{E}\left[|X|^p\right]}{p! c_4^p \nu^p} \leq \sum_{p=0}^{\infty} \frac{(\nu p)^p}{p! c_4^p \nu^p} \leq \sum_{p=0}^{\infty} \left(\frac{e}{c_4}\right)^p = \frac{1}{1 - e/c_4} \leq e$$

provided $c_4 \geq e/(1 - 1/e)$.

[4] The constant e on the righthand side does not have any special meaning and can be replaced by any absolute constant (similar to different scales of a norm.

4 ⇒ 2: Assume $c_4 = 1$. Applying the Markov inequality to $e^{X/\nu}$, it is easy to see that

$$\mathbb{P}\left[X \geq t\right] \leq e^{1-t/\nu}.$$

With the same result for the negative tail, we have

$$\mathbb{P}\left[|X| \geq t\right] \leq \min(2e^{1-t/\nu}, 1) \leq 2\exp\left(-\frac{2t}{5\nu}\right),$$

where in the second inequality we choose a constant c such that both $2e^{1-t/\nu} \leq 2e^{-ct/\nu}$ when t is greater than some threshold and $2e^{-ct/\nu} \geq 1$ when t is greater than the same threshold.

1 ⇒ 3 Using the numerical inequality $|x|^p \leq p^p(e^x + e^{-x})$ for all x and $p > 0$ (**check this!**) with $x = \frac{c_1' X}{\nu}$ and then taking the expectation yields

$$\begin{aligned}\mathbb{E}\left[\left|\frac{c_1' X}{\nu}\right|^p\right] &\leq \mathbb{E}\left[p^p\left(\exp\left(\frac{c_1' X}{\nu}\right) + \exp\left(\frac{-c_1' X}{\nu}\right)\right)\right] \\ &\leq 2p^p\exp\left(c_1\frac{(c_1')^2}{\nu^2}\nu^2\right),\end{aligned}$$

which gives 3 after simplification.

3 ⇒ 1 Assume $c_3 = 1$ for simplicity. By Taylor's expansion we have

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda X\right)\right] &= 1 + \mathbb{E}\left[\lambda X\right] + \sum_{p=2}^{\infty}\frac{\lambda^p\mathbb{E}\left[X^p\right]}{p!} \\ &\leq 1 + \sum_{p=2}^{\infty}\frac{(\lambda p\nu)^p}{p!} \\ &\leq 1 + \sum_{p=2}^{\infty}(\lambda e\nu)^p \qquad (\text{use } p! \geq (p/e)^p) \\ &= 1 + \frac{(\lambda e\nu)^2}{1 - \lambda e\nu} \\ &\leq 1 + 2(\lambda e\nu)^2 \qquad (\text{assume } \lambda e\nu \leq 1/2) \\ &\leq \exp\left(2(\lambda e\nu)^2\right),\end{aligned}$$

which concludes the proof with $c_1 = 2e^2$ and $c_1' = 2e$. ■

1.4 Martingale Methods and Bounded Differences Inequality

Martingale is defined based on conditional expectation, which is about finding an equivalent version (by preserving expectation) of a random variable under given information pool (Wikipedia is a good source for more details). In general, a sequence $\{Z_k\}_{k=0}^{\infty}$ of random variables adapted to a filtration $\{\mathcal{F}_k\}_{k=0}^{\infty}$, the pair $\{(Z_k, \mathcal{F}_k)\}_{k=0}^{\infty}$ is called a martingale if for all $k \geq 0$,

$$\mathbb{E}\left[|Z_k|\right] < \infty, \quad \text{and} \quad \mathbb{E}\left[Z_{k+1}|\mathcal{F}_k\right] = Z_k.$$

However, to avoid the technical difficulties of filtration or σ-algebra, we proceed directly in terms of the marginal difference sequence.

Let $f(x_1, \cdots, x_n) : \mathcal{X}^n \to \mathbb{R}$ be a function and let $X_1, \cdots, X_n$ be independent random variables taking values in the sample space $\mathcal{X}$ (elements of $\mathcal{X}$ can be scalars, vectors, and so on). Define the following martingale difference sequence

$$\begin{aligned} D_1 &= \mathbb{E}\left[f(X_1, \cdots, X_n)\right], \quad D_n = f(X_1, \cdots, X_n), \\ D_k &= \mathbb{E}\left[f(X_1, \cdots, X_n) | X_1, \cdots, X_k\right] - \mathbb{E}\left[f(X_1, \cdots, X_n) | X_1, \cdots, X_{k-1}\right] \\ &= \mathbb{E}_{k+1}[f(X_1, \cdots, X_k, \underbrace{X_{k+1}, \cdots, X_n})] - \mathbb{E}_k[f(X_1, \cdots, X_{k-1}, \underbrace{X_k, X_{k+1}, \cdots, X_n})], \end{aligned} \tag{1.9}$$

where $\mathbb{E}_{k+1}[\cdot]$ means taking expectation with respect to $X_{k+1}, \cdots, X_n$ while keeping $X_1, \cdots, X_n$ unchanged, i.e.,

$$\mathbb{E}_{k+1}\left[f(X_1, \cdots, X_k, X_{k+1}, \cdots, X_n)\right] := \mathbb{E}\left[f(x_1, \cdots, x_k, X_{k+1}, \cdots, X_n)\right]\big|_{x_1 = X_1, \cdots, x_k = X_k}.$$

It is clear that

$$\mathbb{E}\left[D_k | X_1, \cdots, X_{k-1}\right] = 0,$$

and consequently $\mathbb{E}\left[D_k\right] = 0$. Moreover, we have

$$\sum_{k=1}^n D_k = f(X_1, \cdots, X_n) - \mathbb{E}\left[f(X_1, \cdots, X_n)\right],$$

which enables us to study the concentration of $f(X_1, \cdots, X_n)$ around its mean by studying the concentration of the sum $\sum_{k=1}^n D_k$.

Even though $D_k, \ k = 1, \cdots, n$ are not independent to each other, the martingale structure enables us to establish the sub-Gaussian tail once they are bounded.

Theorem 1.34 (Azuma-Hoeffding tail bound) *Let $\{D_k\}_{k=1}^n$ be the martingale difference sequence defined in* (1.9)*. Suppose that $A_k \leq D_k \leq B_k$ almost surely for all $k \geq 1$, where A_k and B_k are functions of $X_1, \cdots, X_{k-1}$. If $B_k - A_k \leq L_k$, then for all $t \geq 0$, we have*

$$\mathbb{P}\left[\left|\sum_{k=1}^n D_k\right| \geq t\right] \leq 2e^{-\frac{2t^2}{\sum_{k=1}^n L_k^2}}$$

Proof: Noting that $\mathbb{E}\left[D_k | X_1, \cdots, X_{k-1}\right] = 0$, repeating the argument in Example 1.13 for a conditional expectation yields that

$$\mathbb{E}\left[e^{\lambda D_k} | X_1, \cdots, X_{k-1}\right] \leq \exp\left(\frac{\lambda^2 (B_k - A_k)^2}{8}\right) \leq \exp\left(\frac{\lambda^2 L_k^2}{8}\right) \tag{1.10}$$

Consequently,

$$\begin{aligned} \mathbb{E}\left[e^{\lambda \sum_{k=1}^n D_k}\right] &= \mathbb{E}\left[\mathbb{E}\left[e^{\lambda \sum_{k=1}^n D_k} | X_1, \cdots, X_{n-1}\right]\right] \\ &= \mathbb{E}\left[e^{\lambda \sum_{k=1}^{n-1} D_k} \mathbb{E}\left[e^{\lambda D_n} | X_1, \cdots, X_{n-1}\right]\right] \end{aligned}$$

$$\leq e^{\lambda^2 L_n^2/8}\mathbb{E}\left[e^{\lambda\sum_{k=1}^{n-1} D_k}\right].$$

Thus, iterating this procedure yields $\mathbb{E}\left[e^{\lambda\sum_{k=1}^{n} D_k}\right] \leq e^{\lambda^2\sum_{k=1}^{n} L_k^2/8}$, which means that $\sum_{k=1}^{n} D_k$ is sub-Gaussian with parameter $\nu^2 = \frac{\sum_{k=1}^{n} L_k^2}{4}$, and an application of the former Hoeffding inequality yields the desired tail bound. ■

Remark 1.35 *There are two key ingredients in the above proof: one is the sub-Gaussian type property but for the conditional expectation; the other one is the tensorization property of the moment generating function but for martingale difference sequence.*

Exercise 1.36 *Write out the details for the proof of* (1.10)*.*

Since Azuma-Hoeffding inequality actually shows the concentration of $f(X_1,\cdots,X_n)$ around its mean with the proviso that D_k are bounded, a natural question will be for which f the corresponding D_k are bounded. Next we are going to show that this is the case if f does not fluctuate with each argument too much, leading to the bounded difference inequality, i.e., the McDiarmid inequality. This result reveals a connection between stability and concentration: if a function $f(x_1,\cdots,x_n)$ is not too sensitive to any of its coordinates x_i, then it is anticipated that $f(X_1,\cdots,X_n)$ ($X_i,\ i=1,\cdots,n$ are independent or weakly independent) is close to its mean. This is also the first concentration result in this course that is beyond the sum of independent random variables, as well as a benchmark concentration inequality we will revisit a few times.

Theorem 1.37 (McDiarmid inequality/Bounded difference inequality) *Let $X_k, k=1,\cdots,n$ be independent random variables taking values in $\mathcal{X}$, where $\mathcal{X}$ is the sample space. Suppose that a function $f:\mathcal{X}^n\to\mathbb{R}$ satisfies the bounded difference property*

$$|f(x_1,\cdots,x_{k-1},x_k,x_{k+1},\cdots,x_n) - f(x_1,\cdots,x_{k-1},x_k',x_{k+1},\cdots,x_n)| \leq L_k$$

with parameters $(L_1,\cdots,L_n)$ for all $x_1,\cdots,x_n,x_k'\in\mathcal{X}$.Then

$$\mathbb{P}\left[|f(X)-\mathbb{E}\left[f(X)\right]|\geq t\right] \leq 2e^{-\frac{2t^2}{\sum_{k=1}^{n} L_k^2}}.$$

Proof: Define D_k as in (1.9). By the last theorem we only need to show D_k is bounded. To this end, define

$$A_k = \inf_{x\in\mathcal{X}}\mathbb{E}_{k+1}\left[f(X_1,\cdots,X_{k-1},x,\underbrace{X_{k+1},\cdots,X_n})\right] - \mathbb{E}_k\left[f(X_1,\cdots,X_{k-1},\underbrace{X_k,X_{k+1},\cdots,X_n})\right]$$

and

$$B_k = \sup_{x\in\mathcal{X}}\mathbb{E}_{k+1}\left[f(X_1,\cdots,X_{k-1},x,\underbrace{X_{k+1},\cdots,X_n})\right] - \mathbb{E}_k\left[f(X_1,\cdots,X_{k-1},\underbrace{X_k,X_{k+1},\cdots,X_n})\right].$$

It is clear that $A_k\leq D_k\leq B_k$ almost surely. Moreover, we have

$$B_k - A_k = \sup_{x\in\mathcal{X}}\mathbb{E}_{k+1}\left[f(X_1,\cdots,X_{k-1},x,\underbrace{X_{k+1},\cdots,X_n})\right] - \inf_{x\in\mathcal{X}}\mathbb{E}_{k+1}\left[f(X_1,\cdots,X_{k-1},x,\underbrace{X_{k+1},\cdots,X_n})\right]$$

$$\le \sup_{x,y\in\mathcal{X}} \left| \mathbb{E}_{k+1}\left[f(X_1,\cdots,X_{k-1},x,\underbrace{X_{k+1},\cdots,X_n}) \right] - \mathbb{E}_{k+1}\left[f(X_1,\cdots,X_{k-1},y,\underbrace{X_{k+1},\cdots,X_n}) \right] \right|$$

$$= \sup_{x,y\in\mathcal{X}} \left| \mathbb{E}_{k+1}\left[f(X_1,\cdots,X_{k-1},x,\underbrace{X_{k+1},\cdots,X_n}) - f(X_1,\cdots,X_{k-1},y,\underbrace{X_{k+1},\cdots,X_n}) \right] \right|$$

$$\le L_k,$$

as desired. ■

Exercise 1.38 *Show how to prove the result in Example 1.17 using the McDiarmid inequality.*

Example 1.39 (Rademacher complexity) *Let $\{\varepsilon_k\}_{k=1}^n$ be an i.i.d sequence of Rademacher variables, namely*

$$\mathbb{P}\left[\varepsilon_k = 1\right] = \mathbb{P}\left[\varepsilon_k = -1\right] = \frac{1}{2},$$

and let $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)$. Given a subset $\mathcal{A}$ of $\mathbb{R}^n$, define the random variable

$$Z = \sup_{a\in\mathcal{A}} \left[\sum_{k=1}^n a_k \varepsilon_k \right] = \sup_{a\in\mathcal{A}} \left[\langle a, \varepsilon \rangle \right].$$

The Rademacher complexity, denoted $\mathcal{R}_n(\mathcal{A})$, is defined as the expectation of Z,

$$\mathcal{R}_n(\mathcal{A}) = \mathbb{E}\left[Z\right].$$

Here the random variable Z and its expectation measures the size of $\mathcal{A}$ based on the Rademacher sequence. Roughly speaking, it measures the "diameter" of the set in different directions randomly and then computes the average. (Again, when it is not clear how to do, do randomly). They also reflect how strong the set $\mathcal{A}$ looks like a random set defined by the Rademacher sequence. For example, if $\mathcal{A} = \{1, -1\}^n$, then it is equal to ε in certain sense.

We want to show that the McDiarmid inequality can be used to establish the concentration of Z. Define

$$f(x_1, \cdots, x_n) = \sup_{a\in\mathcal{A}} \left[\sum_{k=1}^n a_k x_k \right], \quad x_k \in \{1, -1\}.$$

it suffices to show that f satisfies the bounded difference property. To this end, we have

$$\begin{aligned}
& f(x_1, \cdots, x_{k-1}, x_k, x_{k+1}, x_n) - f(x_1, \cdots, x_{k-1}, x_k', x_{k+1}, x_n) \\
&= \sup_{a\in\mathcal{A}} \left[\sum_{k=1}^n a_k x_k \right] - \sup_{a\in\mathcal{A}} \left[\sum_{j=1}^{k-1} a_j x_j + a_k x_k' + \sum_{j=k+1}^n a_j x_j \right] \\
&\le \sup_{a\in\mathcal{A}} \left[\left(\sum_{k=1}^n a_k x_k \right) - \left(\sum_{j=1}^{k-1} a_j x_j + a_k x_k' + \sum_{j=k+1}^n a_j x_j \right) \right] \\
&= \sup_{a\in\mathcal{A}} a_k (x_k - x_k')
\end{aligned}$$

$$\leq 2\sup_{a\in\mathcal{A}}|a_k|,$$

where the last line follows from the fact $x_k,\ x_k' \in \{1,-1\}$. Similarly, we have

$$f(x_1,\cdots,x_{k-1},x_k',x_{k+1},x_n) - f(x_1,\cdots,x_{k-1},x_k,x_{k+1},x_n) \leq 2\sup_{a\in\mathcal{A}}|a_k|.$$

Consequently,

$$|f(x_1,\cdots,x_{k-1},x_k',x_{k+1},x_n) - f(x_1,\cdots,x_{k-1},x_k,x_{k+1},x_n)| \leq 2\sup_{a\in\mathcal{A}}|a_k|.$$

Thus, by the McDiarmid inequality we can see that Z is sub-Gaussian with parameter $\nu^2 = \sum_{k=1}^n \sup_{a\in\mathcal{A}}|a_k|^2$. Later, we will show that this parameter can be sharpened to $\sup_{a\in\mathcal{A}}\sum_{k=1}^n |a_k|^2$. ***To some extend, this has motivated the development of other machinaries for establishing the concentration inequality.***

Example 1.40 *Let $X_k,\ k=1,\cdots,n$ be bounded random vectors in $\mathbb{R}^d$ satisfying $\mathbb{E}[X_k]=0$ and $\|X_k\|_2 \leq B$. We want to study the concentration of*

$$\left\|\frac{1}{n}\sum_{k=1}^n X_k\right\|_2$$

around the mean $\mathbb{E}\left[\left\|\frac{1}{n}\sum_{k=1}^n X_k\right\|_2\right]$. Let $f(x_1,\cdots,x_n) = \left\|\frac{1}{n}\sum_{k=1}^n x_k\right\|_2$, where $x_k \in \mathbb{R}^n$. Then, by triangular inequality

$$|f(x_1,\cdots,x_{k-1},x_k,x_{k+1},\cdots,x_n) - f(x_1,\cdots,x_{k-1},x_k',x_{k+1},\cdots,x_n)| \leq \frac{1}{n}\|x_k - x_k'\|_2 \leq \frac{2B}{n}.$$

Thus, the application of the bounded difference inequality yields that

$$\mathbb{P}\left[\left|\left\|\frac{1}{n}\sum_{k=1}^n X_k\right\|_2 - \mathbb{E}\left[\left\|\frac{1}{n}\sum_{k=1}^n X_k\right\|_2\right]\right| \geq t\right] \leq 2\mathrm{exp}\left(-\frac{nt^2}{2B^2}\right).$$

If we further assume $\mathbb{E}\left[\|X_k\|_2^2\right] \leq \sigma^2$. Then

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{k=1}^n X_k\right\|_2\right] \leq \left(\mathbb{E}\left[\left\|\frac{1}{n}\sum_{k=1}^n X_k\right\|_2^2\right]\right)^{1/2} = \left(\frac{1}{n^2}\sum_{k=1}^n \mathbb{E}\left[\|X_k\|_2^2\right]\right)^{1/2} \leq \frac{\sigma}{\sqrt{n}}.$$

Consequently, we have

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{k=1}^n X_k\right\|_2 \geq \frac{\sigma}{\sqrt{n}} + t\right] \leq 2\mathrm{exp}\left(-\frac{nt^2}{2B^2}\right).$$

Example 1.41 *As pointed out in a motivation example, the analysis of the generalization error in empirical risk minimization eventually boils down to the bound of quantity in the form of*

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

where $X_k, k = 1, \cdots, n$ are random vectors. Suppose $|f|_\infty < B$ for all $f \in \mathcal{F}$. Letting

$$g(x_1, \cdots, x_n) = \sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{k=1}^{n} f(x_k) - \mathbb{E}\left[f(X)\right] \right|,$$

we have

$$\left|g(x_1, \cdots, x_{k-1}, x_k, x_{k+1}, \cdots, x_n) - g(x_1, \cdots, x_{k-1}, x_k', x_{k+1}, \cdots, x_n)\right| \le \sup_{f\in\mathcal{F}} \left| \frac{1}{n} f(x_k) - \frac{1}{n} f(x_k') \right| \le \frac{2B}{n}.$$

Then the application of the bounded difference inequality yields

$$\mathbb{P}\left[\left|\sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{k=1}^{n} f(x_k) - \mathbb{E}\left[f(X)\right] \right| - \mathbb{E}\left[\sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{k=1}^{n} f(x_k) - \mathbb{E}\left[f(X)\right] \right|\right]\right| \ge t\right] \le 2\text{exp}\left(-\frac{nt^2}{2B^2}\right)$$

Remark 1.42 *Note that in Examples 1.39 and 1.41, we still need to compute the mean of the random quantity of interest, which will be another focus of the course.*

Remark 1.43 *The bounded difference inequality is very useful and the next two lectures are essentially about generalizing the bounded different inequality by considering different f and $(X_1, \cdots, X_n)$.*

Reading Materials

[1] Martin J. Wainwright, *High-dimensional statistics – A non-asymptotic viewpoint*, Chapters 2.1 and 2.2.

[2] Roman Vershynin, *High-dimensional probability: An introduction with applications in data science*, Chapters 2.5, 2.6, 2.7 and 2.8.