Lecture 1: Concentration Inequalities for Sums of Independent Random Variables

Fu Ouyang

April 23, 2018

A *concentration inequality* quantifies how a random variable X deviates from its mean μ, i.e. an upper bound for $\mathbb{P}(|X-\mu|>t)$ for all $t>0$. Ideal concentration inequalities satisfy the following properties:

- Be *non-asymptotic* (i.e. hold for all N as opposed to $N \to \infty$)
- Can be applied to a large class of distributions (distribution free)
- Dependence on $\dim(X)$ is explicit

1 Warm-up: Concentration Inequalities for Sums of Independent Rademacher Random Variables

A random variable X has *Rademacher* (*symmetric Bernoulli*) distribution if

$$\mathbb{P}(X=1)=\mathbb{P}(X=-1)=\frac{1}{2}$$

Theorem 1.1 (Hoeffding's inequality). *Let $X_1,...,X_N$ be independent Rademacher random variables, and $a=(a_1,...,a_N)\in\mathbb{R}^N$. Then, for any $t>0$, we have*

$$\mathbb{P}\left(\left|\sum_{i=1}^N a_i X_i\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\|a\|_2^2}\right)$$

Proof of Theorem 1.1. Note that $\cosh(x)=(e^x+e^{-x})/2 \leq e^{x^2/2}$ for all $x\in\mathbb{R}$. Then, by Markov's inequality, we have for all $\lambda>0$,

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^N a_i X_i \geq t\right) &= \mathbb{P}\left(\exp\left(\lambda\sum_{i=1}^N a_i X_i\right) \geq \exp(\lambda t)\right) \leq \exp(-\lambda t)\mathbb{E}\exp\left(\lambda\sum_{i=1}^N a_i X_i\right) \\ &= \exp(-\lambda t)\mathbb{E}\prod_{i=1}^N \exp\left(\lambda a_i X_i\right) \leq \exp(-\lambda t)\prod_{i=1}^N \exp\left(\frac{\lambda^2 a_i^2}{2}\right) \\ &= \exp\left(\frac{\lambda^2\|a\|_2^2}{2}-\lambda t\right)\end{aligned}$$

Hence

$$\mathbb{P}\left(\sum_{i=1}^{N} a_i X_i \geq t\right) \leq \min_{\lambda>0}\left[\exp\left(\frac{\lambda^2\|a\|_2^2}{2} - \lambda t\right)\right] = \exp\left(-\frac{t^2}{2\|a\|_2^2}\right)$$

It follows that

$$\mathbb{P}\left(\left|\sum_{i=1}^{N} a_i X_i\right| \geq t\right) \leq \mathbb{P}\left(\sum_{i=1}^{N} a_i X_i \geq t\right) + \mathbb{P}\left(\sum_{i=1}^{N} a_i X_i \leq -t\right) \leq 2\exp\left(-\frac{t^2}{2\|a\|_2^2}\right)$$

$\square$

Remark. The proof is based on bounding the moment generating function (MGF), which is a quite general method.

Remark. Hoeffding's inequality provides a non-asymptotic bound in that it holds for all fixed N.

Remark. Using the same ideas and Hoeffding's lemma, we can obtain the following extension of Hoeffding's inequality for independent bounded random variables.

Theorem 1.2 (Hoeffding's inequality for bounded random variables)**.** *Let $X_1, ..., X_N$ be independent random variables with $X_i \in [m_i, M_i]$ for every X_i. Then, for any $t > 0$, we have*

$$\mathbb{P}\left(\left|\sum_{i=1}^{N}(X_i - \mathbb{E}X_i)\right| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{N}(M_i - m_i)^2}\right)$$

Remark. Why do we need a probability bound of this type? Consider a case with $m_i = 0$ and $M_i = 1$. Theorem 1.2 implies that

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}(X_i - \mathbb{E}X_i)\right| \geq t\right) \leq 2\exp\left(-2Nt^2\right)$$

Let RHS $= \epsilon_N > 0$. Then

$$t = \sqrt{\frac{1}{2N}\log\frac{2}{\epsilon_N}}$$

Hence, we have $|\frac{1}{N}\sum_{i=1}^{N}(X_i - \mathbb{E}X_i)| \leq \sqrt{(2N)^{-1}\log(2/\epsilon_N)}$ with probability at least $1 - \epsilon_N$. For $\epsilon_N = N^{-\alpha}$, we can say that $|\frac{1}{N}\sum_{i=1}^{N}(X_i - \mathbb{E}X_i)| = O(\sqrt{N^{-1}\log N})$ with probability at least $1 - N^{-\alpha}$. Bounds of this type are favored in learning theory, and are sometimes called *PAC-bounds* (for Probably Approximately Correct).

Hoeffding's inequality gives a Gaussian tail bound. But sometimes it is not sharp. For example, consider the setting of *Poisson limit theorem*, S_N has approximately Poisson distribution. The point is that Hoeffding's inequality (Theorem 1.2) is not sensitive to the magnitude of p_i. The following Chernoff's inequality[1] takes this into account and results in a Poisson tail.

[1]General Chernoff bound can be obtained as follows. Define $\psi_X(\lambda) = \log \mathbb{E}\exp\left[\lambda\left(X - \mathbb{E}X\right)\right]$ for $\lambda \in \mathbb{R}$. For $\lambda > 0$,

$$\mathbb{P}\left(X - \mu \geq t\right) = \mathbb{P}(e^{\lambda(X-\mathbb{E}X)} \geq e^{\lambda t}) \leq e^{-\lambda t}\mathbb{E}e^{\lambda(X-\mathbb{E}X)} = \exp(\psi_X(\lambda) - \lambda t)$$

Then $\mathbb{P}\left(X - \mu \geq t\right) \leq \min_{\lambda>0}\exp(\psi_X(\lambda) - \lambda t)$.

Theorem 1.3 (Chernoff's inequality). *Let $X_1, ..., X_N$ be independent Bernoulli random variables with parameter p_i. Let $S_N \equiv \sum_{i=1}^N X_i$ and $\mu \equiv \mathbb{E} S_N$. Then, for any $t > \mu$, we have*

$$\mathbb{P}(S_N \geq t) \leq e^{-\mu} \left(\frac{e\mu}{t}\right)^t$$

Proof of Theorem 1.3. By Markov's inequality, for all $\lambda > 0$, we have

$$\mathbb{P}(S_N \geq t) \leq e^{-\lambda t} \prod_{i=1}^N \mathbb{E} \exp(\lambda X_i) = e^{-\lambda t} \prod_{i=1}^N (p_i e^\lambda + 1 - p_i) \leq \exp[\mu(e^\lambda - 1) - \lambda t]$$

Hence

$$\mathbb{P}(S_N \geq t) \leq \min_{\lambda > 0} \exp[\mu(e^\lambda - 1) - \lambda t] = e^{-\mu} \left(\frac{e\mu}{t}\right)^t$$

□

Remark. By the Chernoff's inequality, Poisson limit theorem and Slutsky's theorem, we have, for $X \sim \text{Pois}(\lambda)$,

$$\mathbb{P}(X \geq t) \leq e^{-\mu} \left(\frac{e\lambda}{t}\right)^t, t > \lambda$$

$$\mathbb{P}(|X - \lambda| \geq t) \leq 2 \exp\left(-\frac{ct^2}{\lambda}\right), t \in (0, \lambda], c > 0$$

This implies that, for small deviation from the mean λ, Poisson has a tail like $N(\lambda, \lambda)$, while for large deviation (far to the right from λ), Poisson has a heavier tail which decays like $(\lambda/t)^t$.

2 Sub-Gaussian Distributions

So far, we have studied concentration inequalities only for Bernoulli-like (or bounded) random variables. This section extends these results for a wider class of distributions called sub-Gaussian distributions (distributions having tails lighter than Gaussian), which contains Gaussian, Bernoulli and all bounded distributions. Concentration results like Hoeffding's inequality can be proved for all sub-Gaussian distributions.

Proposition 2.1 (Sub-Gaussian properties). Let X be a random variable. Then the following properties are equivalent. The parameters $K_i > 0$ appears in these properties differ from each other by at most an absolute constant factor.

(a) (Tail Behavior) For all $t > 0$,

$$\mathbb{P}(|X| \geq t) \leq 2 \exp\left(-\frac{t^2}{K_1^2}\right)$$

(b) (L^p Norm) For all $p \geq 1$,

$$\|X\|_{L^p} \leq K_2 \sqrt{p}$$

(c) (MGF of X^2)[2] For all λ such that $|\lambda| \leq K_3^{-1}$,

$$\mathbb{E}\exp(\lambda^2 X^2) \leq \exp(K_3^2 \lambda^2)$$

If $\mathbb{E}X = 0$, then properties (a)-(c) are also equivalent to

(d) (MGF of X)[3] For all $\lambda \in \mathbb{R}$,

$$\mathbb{E}\exp(\lambda X) \leq \exp(K_4^2 \lambda^2)$$

Proof of Proposition 2.1. The proof processes in the following steps.

1. $(a) \Rightarrow (b)$: By (a) and $\Gamma(x) \leq x^x$ (see Gamma function and Stirling's approximation),

$$\begin{aligned}\mathbb{E}|X|^p &= \int_0^\infty \mathbb{P}(|X|^p \geq u)du = \int_0^\infty pt^{p-1}\mathbb{P}(|X| \geq t)dt \\ &\leq \int_0^\infty 2pt^{p-1}e^{-t^2/K_1^2}dt = K_1^p p\Gamma(p/2) \leq K_1^p p\,(p/2)^{p/2}\end{aligned}$$

   Then $\|X\|_{L^p} = (\mathbb{E}|X|^p)^{1/p} \leq K_1 p^{1/p}\sqrt{p/2} = K_2\sqrt{p}$ with $K_2 = K_1 p^{1/p}/\sqrt{2}$.

2. $(b) \Rightarrow (c)$: Using the Taylor series expansion, (b), Stirling's approximation and $(1-x)e^{2x} \geq 1$ for $x \in [0, 1/2]$,

$$\begin{aligned}\mathbb{E}\exp(\lambda^2 X^2) &= 1 + \sum_{p=1}^\infty \frac{\lambda^{2p}}{p!}\mathbb{E}(X^{2p}) \leq 1 + \sum_{p=1}^\infty \frac{1}{p!}(\lambda K_2)^{2p}(2p)^p \\ &\leq 1 + \sum_{p=1}^\infty (2eK_2^2\lambda^2)^p = \frac{1}{1 - 2eK_2^2\lambda^2} \leq \exp(4eK_2^2\lambda^2)\end{aligned}$$

   provided that $2eK_2^2\lambda^2 < 1$ and hence $K_3 = 2K_2\sqrt{e}$.

3. $(c) \Rightarrow (a)$: (c) implies that $\mathbb{E}\exp(X^2/C^2) \leq 2$ for some $C = K_3/\sqrt{\log 2}$. Then by Markov's inequality,

$$\mathbb{P}(|X| > t) \leq \mathbb{E}\exp(X^2/C^2 - t^2/C^2) \leq 2\exp(-t^2/K_1^2)$$

   for $K_1 = C$.

[2] If (c) holds for all $\lambda \in \mathbb{R}$, then X is a bounded random variable, i.e. $\|X\|_\infty < \infty$.

[3] $\mathbb{E}X = 0$ is also necessary for (d). If (d) holds, then using Taylor's expansion, we have

$$\lambda\mathbb{E}X \leq \sum_{k=1}^\infty \frac{K_4^{2k}\lambda^{2k}}{k!} - \sum_{k=2}^\infty \frac{\lambda^k}{k!}\mathbb{E}X^k$$

For $\lambda \to 0^+$ and $\lambda \to 0^-$, we have $\mathbb{E}X \leq 0$ and $\mathbb{E}X \geq 0$, respectively.

4. $(c) \Rightarrow (d)$: Note that $e^x \leq x + e^{x^2}$ for all $x \in \mathbb{R}$. Then by (c)

$$\mathbb{E}\exp(\lambda X) \leq \mathbb{E}[\lambda X + \exp(\lambda^2 X^2)] = \mathbb{E}\exp(\lambda^2 X^2) \leq \exp(K_3^2 \lambda^2)$$

provided that $|\lambda| \leq K_3^{-1}$. For $|\lambda| > K_3^{-1}$,

$$\mathbb{E}\exp(\lambda X) \leq \mathbb{E}\exp[K_3^2\lambda^2/2 + X^2/(2K_3^2)] \leq e^{(K_3^2\lambda^2+1)/2} \leq \exp(K_3^2\lambda^2)$$

with $K_4 = K_3$.

5. $(d) \Rightarrow (a)$: Hint: Use the same strategy of proving Hoeffding's inequality.

□

A random variable X is called a *sub-Gaussian random variable* if it satisfies one of the equivalent properties (a)-(c) in Proposition 2.1. The *sub-Gaussian norm*[4] of X, denoted $\|X\|_{\psi_2}$ is defined as

$$\|X\|_{\psi_2} = \inf\{t > 0 : \mathbb{E}\exp(X^2/t^2) \leq 2\}$$

We can restate Proposition 2.1 in terms of the sub-Gaussian norm: for some absolute constant $c, C > 0$, we have

(a) $\mathbb{P}(|X| \geq t) \leq 2\exp(-ct^2/\|X\|_{\psi_2}^2)$ for all $t > 0$.

(b) $\|X\|_{L^p} \leq C\|X\|_{\psi_2}\sqrt{p}$ for all $p \geq 1$.

(c) $\mathbb{E}\exp(X^2/\|X\|_{\psi_2}^2) \leq 2$.

(d) If $\mathbb{E}X = 0$, $\mathbb{E}\exp(\lambda X) \leq \exp(C\lambda^2\|X\|_{\psi_2}^2)$ for all $\lambda \in \mathbb{R}$.

Sums of independent sub-Gaussians satisfies the following *rotaion invariance* property.

Proposition 2.2 (Sums of independent sub-Gaussians). Let $X_1, ..., X_N$ be indepednent sub-Gaussian random variables with $\mathbb{E}X = 0$. Then $\sum_{i=1}^N X_i$ is also a sub-Gaussian random variable, and

$$\left\|\sum_{i=1}^N X_i\right\|_{\psi_2}^2 \leq C\sum_{i=1}^N \|X_i\|_{\psi_2}^2$$

where C is an absolute constant.

[4] It is easy to verify that $\|\cdot\|_{\psi_2}$ is indeed a norm on the space of sub-Gaussian random variables. Here we only prove it satisfies the triangle inequality. For two sub-Gaussian random variables X and Y, we have

$$\exp\left(\frac{X+Y}{t_X+t_Y}\right)^2 = \exp\left(\frac{t_X}{t_X+t_Y}\frac{X}{t_X} + \frac{t_Y}{t_X+t_Y}\frac{Y}{t_Y}\right)^2 \leq \frac{t_X}{t_X+t_Y}\exp\left(\frac{X}{t_X}\right)^2 + \frac{t_Y}{t_X+t_Y}\exp\left(\frac{Y}{t_Y}\right)^2$$

where $t_X \equiv \|X\|_{\psi_2}, t_Y \equiv \|Y\|_{\psi_2}$. Taking expectation on both sides gives

$$\mathbb{E}\exp\left(\frac{X+Y}{t_X+t_Y}\right)^2 \leq 2 \Rightarrow \|X+Y\|_{\psi_2} \leq \|X\|_{\psi_2} + \|Y\|_{\psi_2}$$

Proof of Proposition 2.2. [5] By proposition 2.1(d), we have

$$\mathbb{E}\exp(\lambda\sum_{i=1}^{N}X_i) = \prod_{i=1}^{N}\mathbb{E}\exp(\lambda X_i) \le \prod_{i=1}^{N}\exp(C\lambda^2\,\|X_i\|_{\psi_2}^2) = \exp(K^2\lambda^2)$$

where $K^2 = C\sum_{i=1}^{N}\|X_i\|_{\psi_2}^2$. Hence, $\sum_{i=1}^{N}X_i$ is also sub-Gaussian and $\|\sum_{i=1}^{N}X_i\|_{\psi_2} \le C_1K$ for some absolute constant. □

By Proposition 2.1(a), we have the following Hoeffding's inequality for sub-Gaussians.

Theorem 2.3 (General Hoeffding's inequality). *Let $X_1, ..., X_N$ be independent sub-Gaussian random variables with $\mathbb{E}X = 0$. For all $a = (a_1, ..., a_N) \in \mathbb{R}^N$ and $t \ge 0$, we have*

$$\mathbb{P}\left(\left|\sum_{i=1}^{N}X_i\right| \ge t\right) \le 2\exp\left(-\frac{ct^2}{\sum_{i=1}^{N}\|X_i\|_{\psi_2}^2}\right)$$

$$\mathbb{P}\left(\left|\sum_{i=1}^{N}a_iX_i\right| \ge t\right) \le 2\exp\left(-\frac{ct^2}{K^2\|a\|_2^2}\right)$$

where $K = \max_i\|X_i\|_{\psi_2}$.

A random vector $X \in \mathbb{R}^d$ is said to be sub-Gaussian if v^TX is sub-Gaussian for any unit vector $v \in \mathcal{S}^{d-1}$ (unit sphere in $\mathbb{R}^d$). Theorem 2.3 implies that a vector $X = (X_1, ..., X_N)$ of independent sub-Gaussian random variables is also sub-Gaussian.
In results like Hoeffding's inequality among others, we often assume random variables X_i having zero means. If this is not the case, we can always center X_i by subtracting its mean. The following lemma guarantees that the centering does not harm the sub-Gaussian property.

Lemma 2.4 (Centering inequality). *If X is a sub-Gaussian random variable then $X - \mathbb{E}X$ is sub-Gaussian, too, and*

$$\|X - \mathbb{E}X\|_{\psi_2} \le C\|X\|_{\psi_2}$$

where C is an absolute constant.

Proof of Lemma 2.4. It follows by Proposition 2.1(d) that $X - \mathbb{E}X$ is sub-Gaussian if X is a sub-Gaussian. Note that by Proposition 2.1(b)

$$\|\mathbb{E}X\|_{\psi_2} \lesssim |\mathbb{E}X| \le \mathbb{E}|X| = \|X\|_1 \lesssim \|X\|_{\psi_2}$$

[5] Here is an alternative way to prove Proposition 2.2. By Hölder's inequality and Proposition 2.1(d), we have

$$\mathbb{E}\exp\left[\lambda(X+Y)\right] \le (\mathbb{E}e^{\lambda pX})^{1/p}(\mathbb{E}e^{\lambda qY})^{1/q} \le \exp(C\lambda^2p\|X\|_{\psi_2}^2 + C\lambda^2q\|Y\|_{\psi_2}^2)$$

with $p^{-1} + q^{-1} = 1$, $p > 1, q > 1$. Minimizing over (p, q) yields

$$\mathbb{E}\exp\left[\lambda(X+Y)\right] \le \exp[C\lambda^2(\|X\|_{\psi_2} + \|Y\|_{\psi_2})^2]$$

This bound is sharp as Hölder's inequality is sharp.

Then, it follows by triangle inequality that

$$\|X - \mathbb{E}X\|_{\psi_2} \leq \|X\|_{\psi_2} + \|\mathbb{E}X\|_{\psi_2} \leq C\|X\|_{\psi_2}$$

□

Remark. Lemma 2.4 implies that the variance of a sub-Gaussian random variable is finite. More precisely, $\mathbb{V}(X) \lesssim \|X\|_{\psi_2}^2$. To see this, note that by Taylor's expansion, Proposition 2.1(d) and Lemma 2.4, we have

$$\mathbb{E}\exp[\lambda\,(X - \mathbb{E}X)] = 1 + \frac{\lambda^2}{2}\mathbb{V}(X) + o(\lambda^2) \leq 1 + C\lambda^2\|X\|_{\psi_2}^2 + o(\lambda^2)$$

Then, the result follows by dividing λ^2 on both sides and letting $\lambda \to 0$.

3 Sub-Exponential Distributions

In this section, we study the class of distributions that have at least an exponential tail decay, which is heavier than sub-Gaussian.

Proposition 3.1 (Sub-exponential properties). Let X be a random variable. The following properties are equivalent. The parameters $K_i > 0$ appearing in these properties differ from each other by at most an absolute constant factor.

(a) (Tails of X) For all $t \geq 0$,

$$\mathbb{P}(|X| \geq t) \leq 2\exp(-t/K_1)$$

(b) (Moments of X) For all $p \geq 1$,

$$\|X\|_{L^p} \leq K_2 p$$

(c) (MGF of $|X|$)[6] For all λ such that $0 \leq \lambda \leq K_3^{-1}$,

$$\mathbb{E}\exp(\lambda|X|) \leq \exp(K_3\lambda)$$

Moreover, if $\mathbb{E}X = 0$ then properties (a) - (c) are also equivalent to the following one

(d) (MGF of X) For all λ such that $|\lambda| \leq K_4^{-1}$,

$$\mathbb{E}\exp(\lambda X) \leq \exp(K_4^2\lambda^2)$$

Proof of Proposition 3.1. The proof processes in the following steps.

[6] Not like Proposition 2.1(c), this bound cannot be extended for all λ such that $|\lambda| \leq K_3^{-1}$.

1. $(a) \Rightarrow (b)$: By (a) and Stirling's approximation, we have for $p \geq 1$,

$$\mathbb{E}|X|^p = \int_0^\infty \mathbb{P}(|X|^p \geq u)du = p\int_0^\infty t^{p-1}\mathbb{P}(|X|^p \geq t^p)dt \leq 2p\int_0^\infty t^{p-1}\exp(-t/K_1)dt$$
$$= 2K_1^p p! \leq 2K_1^p p^p \Rightarrow \|X\|_{L^p} \leq K_2 p$$

where $K_2 = \sqrt[p]{2}K_1$.

2. $(b) \Rightarrow (c)$: Using Taylor's expansion, (b) and Stirling's approximation, we have

$$\mathbb{E}\exp(\lambda|X|) = \mathbb{E}\sum_{k=0}^\infty \lambda^k \frac{|X|^k}{k!} \leq \sum_{k=0}^\infty \frac{\lambda^k K_2^k p^k}{k!} \leq \sum_{k=0}^\infty (e\lambda K_2 p)^k = \frac{1}{1-eK_2p\lambda}$$

provided $0 \leq \lambda < (eK_2p)^{-1}$. Recall that $e^{2x} \geq (1-x)^{-1}$ for $x \in [0, 1/2]$. Then, it follows that

$$\mathbb{E}\exp(\lambda|X|) \leq \frac{1}{1-eK_2p\lambda} \leq \exp(2eK_2p\lambda)$$

Hence, we can choose $K_3 = 2eK_2p$. Note that when $\lambda = C^{-1} = K_3^{-1}\log 2 \in [0, K_3^{-1}]$, $\mathbb{E}\exp(|X|/C) \leq 2$.

3. $(c) \Rightarrow (a)$: It is easy to show this using Markov's inequality.

4. $(b) \Rightarrow (d)$: Using Taylor's expansion, $\mathbb{E}X = 0$ and (b), we have

$$\mathbb{E}\exp(\lambda X) = 1 + \sum_{k=2}^\infty \frac{\lambda^k \mathbb{E}X^k}{k!} \leq 1 + \sum_{k=2}^\infty \frac{\lambda^k K_2^k k^k}{k!}$$

Then, it follows by Stirling's approximation that

$$\mathbb{E}\exp(\lambda X) \leq 1 + \sum_{k=2}^\infty \frac{\lambda^k K_2^k k^k}{k!} \leq 1 + \sum_{k=2}^\infty \frac{\lambda^k K_2^k k^k}{k^k e^{-k}} = 1 + \sum_{k=2}^\infty \lambda^k e^k K_2^k = 1 + \frac{(\lambda e K_2)^2}{1-\lambda e K_2}$$

provided that $|\lambda e K_2| \leq 1$. Moreover, if $|\lambda e K_2| \leq 1/2$, we have $\mathbb{E}\exp(\lambda X) \leq 1 + 2(\lambda e K_2)^2 \leq \exp(2e^2K_2^2\lambda^2)$. This yields (d) with $K_4 = 2eK_2$.

5. $(d) \Rightarrow (b)$: Note that $|X|^p/p^p \leq e^x + e^{-x}$ for all $x \in \mathbb{R}$ and $p > 0$. Hence, by (d), we have

$$\|X\|_{L^p} = (\mathbb{E}|X|^p)^{1/p} \leq p(\mathbb{E}e^X + \mathbb{E}e^{-X})^{1/p} \leq p(e^{K_4^2} + e^{-K_4^2})^{1/p}$$

provided that $K_4 \geq 1$. This yields (b) with $K_2 = (e^{K_4^2} + e^{-K_4^2})^{1/p}$.

□

A random variable X is called a *sub-exponential random variable* if it satisfies one of the equivalent properties (a)-(c) in Proposition 3.1. The *sub-exponential norm* of X, denoted $\|X\|_{\psi_1}$ is defined as

$$\|X\|_{\psi_1} = \inf\{t > 0 : \mathbb{E}\exp(|X|/t) \leq 2\}$$

It is obvious that any sub-Gaussian distribution is sub-exponential. Furthermore, by the definition of the sub-Gaussian and sub-exponential norms, we have the following relationship between sub-Gaussian and sub-exponential random variables.

Lemma 3.2 (Sub-exponential is sub-Gaussian squared). *A random variable X is sub-Gaussian if and only if X^2 is sub-exponential. Moreover,*

$$\|X^2\|_{\psi_1} = \|X\|_{\psi_2}^2$$

In fact, the result above can be more general.

Lemma 3.3 (Product of sub-Gaussians is sub-exponential). *Let X and Y be sub-Gaussian random variables. Then XY is sub-exponential. Moreover,*

$$\|XY\|_{\psi_1} \leq \|X\|_{\psi_2}\|Y\|_{\psi_2}$$

Proof of Lemma 3.3. W.L.O.G. assume $\|X\|_{\psi_2} = \|Y\|_{\psi_2} = 1$. By definition, we need to show $\mathbb{E}\exp(|XY|) \leq 2$. Using Young's inequality twice yields

$$\mathbb{E}\exp(|XY|) \leq \mathbb{E}\exp(\frac{X^2+Y^2}{2}) \leq \frac{\mathbb{E}\exp(X^2) + \mathbb{E}\exp(Y^2)}{2} \leq 2$$

which completes the proof. □

The following centering inequality is an analog of Lemma 2.4.

Lemma 3.4 (Centering inequality). *Let X be a sub-exponential random variables. Then, for some absolute constant $C > 0$,*

$$\|X - \mathbb{E}X\|_{\psi_1} \leq C\|X\|_{\psi_1}$$

4 Bernstein's Inequality

The following concentration inequalities are for sums of independent sub-exponential random variables[7].

Theorem 4.1 (Bernstein's inequality). *Let $X_1, ..., X_N$ be independent sub-exponential random variables with $\mathbb{E}X = 0$, and $a = (a_1, ..., a_N) \in \mathbb{R}^N$. Then, for any $t \geq 0$,*

$$\mathbb{P}\left(\left|\sum_{i=1}^N X_i\right| \geq t\right) \leq 2\exp\left\{-c\left(\frac{t^2}{\sum_{i=1}^N \|X_i\|_{\psi_1}^2} \wedge \frac{t}{K}\right)\right\}$$

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^N X_i\right| \geq t\right) \leq 2\exp\left\{-c\left(\frac{t^2}{K^2} \wedge \frac{t}{K}\right)N\right\}$$

$$\mathbb{P}\left(\left|\sum_{i=1}^N a_i X_i\right| \geq t\right) \leq 2\exp\left\{-c\left(\frac{t^2}{K^2\|a\|_2^2} \wedge \frac{t}{K\|a\|_\infty}\right)\right\}$$

where $K = \max_i \|X_i\|_{\psi_1}$ and $c > 0$ is an absolute constant.

[7]There exist bounds sharper than Bernstein's inequality for specific distributions, e.g. *Laurent-Massart inequality* for χ^2 distribution.

Proof of Theorem 4.1. Here we only prove the first inequality. By Proposition 3.1, for some absolute constant $c_1 > 0$ and small λ such that $0 < \lambda \le c_1 / \max_i \|X_i\|_{\psi_1}$, we have

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^{N} X_i \ge t\right) &\le e^{-\lambda t} \prod_{i=1}^{N} \mathbb{E}\exp\left(\lambda X_i\right) \le e^{-\lambda t} \prod_{i=1}^{N} \exp(\|X_i\|_{\psi_1}^2 \lambda^2 / c_1^2) \\ &= \exp\left(\lambda^2 \sum_{i=1}^{N} \|X_i\|_{\psi_1}^2 / c_1^2 - \lambda t\right)\end{aligned}$$

To tighten the bound, we solve

$$\lambda^* = \arg\min_{\lambda > 0} \lambda^2 \sum_{i=1}^{N} \|X_i\|_{\psi_1}^2 / c_1^2 - \lambda t \text{ s.t. } 0 < \lambda \le \frac{c_1}{\max_i \|X_i\|_{\psi_1}}$$

If $c_1^2 t / (2 \sum_{i=1}^{N} \|X_i\|_{\psi_1}^2) \le c_1 / \max_i \|X_i\|_{\psi_1}$,

$$\lambda^* = \frac{c_1^2 t}{2 \sum_{i=1}^{N} \|X_i\|_{\psi_1}^2}, \mathbb{P}\left(\sum_{i=1}^{N} X_i \ge t\right) \le \exp\left(-\frac{c_1 t^2}{4 \sum_{i=1}^{N} \|X_i\|_{\psi_1}^2}\right)$$

otherwise,

$$\lambda^* = \frac{c_1}{\max_i \|X_i\|_{\psi_1}}, \mathbb{P}\left(\sum_{i=1}^{N} X_i \ge t\right) \le \exp\left(-\frac{c_1 t}{2 \max_i \|X_i\|_{\psi_1}}\right)$$

Then, the desired bound follows. □

When X_i are all bounded random variables, the Bernstein's inequality can be strengthened.

Theorem 4.2 (Bernstein's inequality for bounded distributions)**.** *Let $X_1, ..., X_N$ be independent sub-exponential random variables with $\mathbb{E}X = 0$, such that $|X_i| \le K$ for all $1 \le i \le N$. Then, for any $t \ge 0$,*

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

where $\sigma^2 = \mathbb{V}(\sum_{i=1}^{N} X_i) = \sum_{i=1}^{N} \mathbb{E}X_i^2$.

Proof of Theorem 4.2. First, note that by induction

$$e^x = 1 + x + \frac{x^2}{2} \sum_{k=1}^{\infty} \frac{x^k}{(2+k)!/2} \le 1 + x + \frac{x^2}{2} \sum_{k=1}^{\infty} \frac{x^k}{3^k} \le 1 + x + \frac{x^2/2}{1 - |x|/3}$$

when $|x| < 3$. Then for a random variable X with $|X| \le K$ and $|\lambda| < 3/K$, we have

$$\mathbb{E}\exp(\lambda X) \le \mathbb{E}\left(1 + \lambda X + \frac{\lambda^2 X^2/2}{1 - |\lambda X|/3}\right) \le 1 + \frac{\lambda^2 \mathbb{E}X^2/2}{1 - |\lambda| K/3} \le \exp\left(\frac{\lambda^2 \mathbb{E}X^2/2}{1 - |\lambda| K/3}\right)$$

With this bound for the MGF of X, the proof can be done very similarly as Theorem 4.1. □

5 Orlicz Spaces

Sub-Gaussian and sub-exponential distributions can be introduced within a general framework of *Orlicz spaces*. A function $\psi : [0, \infty) \to [0, \infty)$ is called an *Orlicz function* if ψ is convex, increasing, and satisfies $\psi(0) = 0$ and $\psi(x) \to \infty$ as $x \to \infty$. The *Orlicz norm* of a random variable X for ψ is defined as

$$\|X\|_\psi = \inf\{t > 0 : \mathbb{E}\psi(|X|/t) \le 1\}$$

The *Orlicz space* $L_\psi = L_\psi(\Omega, \mathcal{A}, \mathbb{P})$ consists of all random variables X on the probability space with finite Orlicz norm, i.e.

$$L_\psi = \{X : \|X\|_\psi < \infty\}$$

- When $\psi(x) = x^p$ for $p \ge 1$, Orlicz norm is the L^p norm $\|\cdot\|_{L^p}$.
- When $\psi(x) = e^{x^2} - 1$, Orlicz norm is the sub-Gaussian norm $\|\cdot\|_{\psi_2}$.
- When $\psi(x) = e^x - 1$, Orlicz norm is the sub-exponential norm $\|\cdot\|_{\psi_1}$.

It is easy to verify that for all $p \in [1, \infty)$, we have

$$L^\infty \subset L_{\psi_2} \subset L_{\psi_1} \subset L^p$$

The following concentration inequalities are for sums of independent sub-exponential random variables.

6 Other Concentration Inequalities

The following *bounded differences inequality*, also called *McDiarmid's inequality*, can be thought of as a generalization of Hoeffding's inequality.

Theorem 6.1 (Bounded differences inequality). *Let $X_1, ..., X_N$ be independent random variables, $f : \mathbb{R}^N \to \mathbb{R}$ be a meaurable function and $X = (X_1, ..., X_N)$. Assume that for any index i and any $X'_{(i)} = (X_1, ..., X_{i-1}, X'_i, X_{i+1}, ..., X_N)$, there is an absolute constant $c_i > 0$ such that $|f(X) - f(X'_{(i)})| \le c_i$. Then, for any $t > 0$, we have*

$$\mathbb{P}(|f(X) - \mathbb{E}f(X)| \ge t) \le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^N c_i^2}\right)$$

The *Bennett's inequality* below generalizes Chernoff's bound.

Theorem 6.2 (Bennett's inequality). *Let $X_1, ..., X_N$ be independent random variables. Assume that $|X_i - \mathbb{E}X_i| \le K$ a.s. for every i. Then, for any $t > 0$, we have*

$$\mathbb{P}\left(\left|\sum_{i=1}^N (X_i - \mathbb{E}X_i)\right| \ge t\right) \le 2\exp\left(-\frac{\sigma^2}{K^2} h\left(\frac{Kt}{\sigma^2}\right)\right)$$

where $\sigma^2 = \sum_{i=1}^N \mathbb{V}(X_i)$ and $h(x) = (1+x)\log(1+x) - x$.

Remark. Note that for small Kt/σ^2, $h(Kt/\sigma^2) \lesssim t^2$, then Bennett's inequality gives approximately the Gaussian tail bound, while for large Kt/σ^2, $h(Kt/\sigma^2) \gtrsim t \log t$, which leads to a Poisson tail bound.

Remark. Both McDiarmid's inequality and Bennett's inequality can be proved by the same general method as Hoeffding's inequality or Chernoff's inequality.

Problem Set

1. [V] Exercise 2.2.9
2. [V] Exercise 2.3.5
3. [V] Exercise 2.5.10
4. [V] Exercise 2.5.11
5. [V] Exercise 2.6.5 - 2.6.7
6. [V] Exercise 2.6.9
7. [V] Exercise 2.7.3
8. Reading assignment: [V] Section 2.4 for an application of Chernoff's inequality

Appendix

- The following classic inequalities will be very useful.
  - *Jensen's inequality*: For any random variable X and a convex function $\varphi : \mathbb{R} \to \mathbb{R}$,
  $$\varphi(\mathbb{E}X) \le \mathbb{E}\varphi(X)$$
  which implies $\|X\|_{L^p} \le \|X\|_{L^q}$ for any $0 \le p \le q \le \infty$.
  - *Minkowski's inequality*: For any $p \ge 1$ and any random variables $X, Y \in L^p$,
  $$\|X + Y\|_{L^p} \le \|X\|_{L^p} + \|X\|_{L^p}$$
  - *Hölder's inequality*: For $p > 1$ and $q > 1$ such that $p^{-1} + q^{-1} = 1$, if $X \in L^p$ and $Y \in L^q$, $\|XY\|_{L^1} \le \|X\|_{L^p} \|Y\|_{L^q}$.
  - *Markov's inequality*: If ϕ is a monotonically increasing nonnegative function, X is a random variable, $t > 0$ and $\phi(t) > 0$, then
  $$\mathbb{P}\left(|X - \mathbb{E}X| \ge t\right) \le \frac{\mathbb{E}\varphi\left(|X - \mathbb{E}X|\right)}{\varphi(t)}$$

An immediate corollary is

$$\mathbb{P}\left(|X-\mathbb{E}X| \geq t\right) \leq \min_{k=1,2,\ldots} \frac{\mathbb{E}|X-\mathbb{E}X|^k}{t^k}$$

This is a good bound if we know all moments of X, but this is in general unrealistic.

- Numeric inequalities:
  - For all $x \in \mathbb{R}$, $\cosh(x) = (e^x + e^{-x})/2 \leq e^{x^2/2}$.
  - For all $x \in \mathbb{R}$, $1 + x \leq e^x$.
  - For $x \in [0, 1/2]$, $(1-x)e^{2x} \geq 1$.
  - For all $x \in \mathbb{R}$, $e^x \leq x + e^{x^2}$.
  - For all $x \in \mathbb{R}$ and $p > 0$, $|X|^p/p^p \leq e^x + e^{-x}$.
- *Gamma function*:

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt = 2\int_0^\infty t^{2x-1}e^{-t^2}dt$$

- *Stirling's Approximation*:

$$n! \approx n^n e^{-n}\sqrt{(2n+\frac{1}{3})\pi}$$

- *Integral identity*:
  - Let X be a non-negative random variable.

$$\mathbb{E}X = \int_0^\infty \mathbb{P}(X > t)dt$$

    To see this, note that

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

  - More generally, for any random variable X

$$\mathbb{E}X = \int_0^\infty \mathbb{P}(X > t)dt - \int_{-\infty}^0 \mathbb{P}(X < t)dt$$

  - For $p \in (0, \infty)$

$$\mathbb{E}|X|^p = \int_0^\infty pt^{p-1}\mathbb{P}(|X| > t)dt$$

- *Berry-Esseen CLT*: Let $X_1, X_2, ...$ be a sequence of iid random variables with mean μ and variance σ^2. Define

$$S_N = \sum_{i=1}^{N} X_i, Z_N = \frac{S_N - \mathbb{E}S_N}{\sqrt{V(S_N)}}$$

For every $t \in \mathbb{R}$, we have

$$|\mathbb{P}(Z_N \geq t) - \mathbb{P}(Z \geq t)| \leq \frac{\mathbb{E}|X_1 - \mu|^3}{\sigma^3 \sqrt{N}}$$

where $Z \sim N(0,1)$. This implies that the approximation error of the CLT is $O(N^{-1/2})$. Furthermore, this is a sharp bound (consider $X_i \overset{iid}{\sim} Bernoulli(1/2)$, $\mathbb{P}(S_N = N/2) - \mathbb{P}(Z = N/2) \approx N^{-1/2}$).

- *Hoeffding's lemma*: Let X be a random variable with $a \leq X \leq b$. Then for all $\lambda \in \mathbb{R}$,

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

To show this, let $Y = X - \mathbb{E}X$. Note that

$$\mathbb{E}e^{\lambda Y} \leq -\frac{a}{b-a}e^{tb} + \frac{b}{b-a}e^{ta} = e^{g(u)}, u = \lambda\left(b-a\right)$$

where

$$g(u) = \log(1 - \theta + \theta e^u) - \theta u$$
$$u = \lambda(b-a), \theta = -\frac{a}{b-a} < 0$$

It is easy to verify that $g\left(0\right) = g'\left(0\right) = 0$, $g''\left(u\right) \leq 1/4$ for all $u > 0$. Then by Taylor's expansion and mean value theorem, there exists a $\xi \in (0, u)$ such that

$$g\left(u\right) = g\left(0\right) + ug'\left(0\right) + \frac{u^2}{2}g''\left(\xi\right) \leq \frac{u^2}{8}$$

The following proof gives a slightly weaker version of Hoeffding's lemma. But the proof itself is well worth reading, as it uses a very useful technique in probability theory known as *symmetrization*.

Let $\mathbb{E}_Z$ indicate expectations taken with respect to a random vector Z, X' be an iid copy of X, and ϵ be a Rademacher random variable. By Jensen's inequality, we have

$$\mathbb{E}\exp[\lambda(X - \mathbb{E}X)] = \mathbb{E}_X \exp[\lambda(X - \mathbb{E}_{X'}X')] \leq \mathbb{E}_{X,X'} \exp[\lambda(X - X')]$$

Note that $X - X' \overset{d}{=} \epsilon(X - X')$, $|X - X'| \leq b - a$, and $e^x + e^{-x} \leq 2e^{x^2/2}$ for all $x \in \mathbb{R}$. Then, we have

$$\begin{aligned}\mathbb{E}_{X,X'} \exp[\lambda\epsilon(X - X')] &= \mathbb{E}_{X,X'}\mathbb{E}_{\epsilon}\{\exp[\lambda\epsilon(X - X')]|X, X'\} \\ &\leq \mathbb{E}_{X,X'} \exp\left(\frac{\lambda^2\left|X - X'\right|^2}{2}\right) \leq \exp\left(\frac{\lambda^2\left(b-a\right)^2}{2}\right)\end{aligned}$$

- *Mills inequality* (tails of normal distributions): Let $X \sim N(0,1)$. For all $t > 0$,

$$\mathbb{P}(X > t) = \int_t^\infty \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}) dx \leq \frac{1}{t\sqrt{2\pi}} \int_t^\infty x \exp(-\frac{x^2}{2}) dx = \frac{1}{t\sqrt{2\pi}} \exp(-\frac{t^2}{2})$$