Notes on Statistical Learning Theory

Junming Yin

This version: June 13, 2021

Contents

Chapter 1

Concentration Inequalities

1.1 Preliminaries: o_P and O_P notation

Definition 1 (o_P)**.** Let $\{X_n\}$ be a sequence of random variables, we say $X_n = o_P(1)$ if $X_n \xrightarrow{p} 0$. That is, for every $\epsilon > 0$,

$$\lim_n \mathbb{P}(|X_n| > \epsilon) = 0.$$

More generally, $X_n = o_P(Y_n)$ means $X_n/Y_n = o_P(1)$.

Definition 2 (O_P)**.** Let $\{X_n\}$ be a sequence of random variables, we say $X_n = O_P(1)$ if for every $\epsilon > 0$, there exists $M(\epsilon) > 0$ such that

$$\mathbb{P}(|X_n| > M(\epsilon)) < \epsilon, \quad \text{for all } n.$$

Thus, there exists a compact set to which all X_n give probability "almost" one. $\{X_n\}$ is also called *uniformly tight* and *bounded in probability.* More generally, $X_n = O_P(Y_n)$ means $X_n/Y_n = O_P(1)$.

Remark. *Sometimes the above definition applies to only sufficiently large n. That is for every $\epsilon > 0$, there exists $M(\epsilon)$ and $n_0(\epsilon)$ such that the above inequality holds for all $n > n_0(\epsilon)$.*

Proposition 1.

(a) $X_n = o_P(1)$ implies $X_n = O_P(1)$. Every converging random sequence is bounded in probability.

(b) (Prohorov) If $X_n \xrightarrow{d} X$ for some X, then $X_n = O_P(1)$. Every weakly converging random sequence is uniformly tight.

Proof. (a) By definition, for every $\epsilon > 0$, there exists an $n_0(\epsilon)$ such that for all $n \geq n_0(\epsilon)$,

$$\mathbb{P}(|X_n| > 1) < \epsilon.$$

Next, we pick $M_0(\epsilon)$ so that $\mathbb{P}(|X_i| > M_0(\epsilon)) < \epsilon$ for $i = 1, \dots, n_0(\epsilon) - 1$. Thus, we have for $M(\epsilon) = \max\{M_0(\epsilon), 1\}$ that

$$\mathbb{P}(|X_n| > M(\epsilon)) < \epsilon, \quad \text{for all } n,$$

which proves $X_n = O_P(1)$.

(b) For every $\epsilon > 0$, we pick $M_1(\epsilon)$ to be the continuity point of distribution function of X so that

$$\mathbb{P}\left(|X| > M_1(\epsilon)\right) < \epsilon/2.$$

By definition of weak convergence, $\mathbb{P}\left(|X_n| > M_1(\epsilon)\right) \to \mathbb{P}\left(|X| > M_1(\epsilon)\right)$ as $n \to \infty$. Thus there exists an $n_0(\epsilon)$ such that for all $n \geq n_0(\epsilon)$,

$$\mathbb{P}\left(|X_n| > M_1(\epsilon)\right) < \mathbb{P}\left(|X| > M_1(\epsilon)\right) + \epsilon/2 < \epsilon.$$

Next, we pick $M_0(\epsilon)$ so that $\mathbb{P}(|X_i| > M_0(\epsilon)) < \epsilon$ for $i = 1, \dots, n_0(\epsilon) - 1$. Thus, for $M(\epsilon) = \max\{M_0(\epsilon), M_1(\epsilon)\}$, we have

$$\mathbb{P}(|X_n| > M(\epsilon)) < \epsilon, \quad \text{for all } n,$$

which completes the proof that $X_n = O_P(1)$.

□

Remark. *We may use part(b) to prove (a) directly: $X_n = o_P(1)$ implies $X_n \xrightarrow{d} 0$, which immediately shows $X_n = O_P(1)$ by part (b). Nevertheless, the condition in part (b) is not weaker than the one in part (a). Part (a) requires X_n to converge to a constant 0 in probability whereas part (b) requires X_n to converge to a random variable X (not necessarily 0) in distribution.*

Example 1.

(a) By weak law of large numbers, $\bar{X}_n \xrightarrow{p} \mu$ so that $\bar{X}_n - \mu = o_P(1)$. Here, $\bar{X}_n = \sum_{i=1}^n X_i/n$ is the sample mean of i.i.d. integrable random variables with expected value $\mathbb{E}[X_1] = \mu$.

(b) By central limit theorem, $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ so that $\bar{X}_n - \mu = O_P\left(\frac{1}{\sqrt{n}}\right)$. Here, σ^2 is the variance of i.i.d. random variables X_i.

Proposition 2.

(a) $X_n = o_P(a_n)$ and $Y_n = o_P(b_n)$, then $X_nY_n = o_P(a_nb_n)$.

(b) *$X_n = o_P(a_n)$ and $Y_n = o_P(b_n)$, then $X_n + Y_n = o_P(\max\{a_n, b_n\})$.*

(c) *$X_n = O_P(a_n)$ and $Y_n = O_P(b_n)$, then $X_n Y_n = O_P(a_n b_n)$.*

(d) *$X_n = O_P(a_n)$ and $Y_n = O_P(b_n)$, then $X_n + Y_n = O_P(\max\{a_n, b_n\})$.*

(e) *$X_n = o_P(a_n)$ and $Y_n = O_P(b_n)$, then $X_n Y_n = o_P(a_n b_n)$.*

(f) *$X_n = o_P(a_n)$ and $Y_n = O_P(b_n)$, then $X_n + Y_n = O_P(\max\{a_n, b_n\})$.*

Proof. TODO □

Remark. *In multiplication, o_P always wins; while in addition, O_P always wins.*

TODO: Delta method, Slutsky, Continuous mapping theorem, Lehmann (2004, Theorem 2.3.4)

1.2 Motivation

Concentration inequalities are concerned bounding random fluctuations of functions of many independent random variables. One key property of those inequalities is that the random variables are only required to be *independent*, but not necessarily *identically distributed*. One general form is as follows:

$$\mathbb{P}\left(\left|g(Z_1, Z_2, \ldots, Z_n) - \mathbb{E}[g(Z_1, Z_2, \ldots, Z_n)]\right| \geq \epsilon\right) \leq \delta_n,$$

where $Z_1, Z_2, \ldots, Z_n$ are independent random variables and $g : \mathcal{Z}^n \to \mathbb{R}$ is a real-valued measurable function. We'd like to have $\delta_n \to 0$ as $n \to \infty$. More generally, we may need *uniform bounds* of the above form:

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}}\left|g(Z_1, Z_2, \ldots, Z_n) - \mathbb{E}[g(Z_1, Z_2, \ldots, Z_n)]\right| \geq \epsilon\right) \leq \delta_n \tag{1.1}$$

over a function class $\mathcal{G}$.

Example 2. Consider the empirical risk minimization framework for binary classification problems. Given a decision rule $f : \mathcal{X} \to \{0, 1\}$, the population risk is defined as

$$R(f) = \mathbb{P}\left(Y \neq f(X)\right),$$

and the empirical risk on the training data $\mathcal{D}_n = \{(X_i, Y_i), i = 1 \cdots n\}$ is

$$\widehat{R}_n(f) = \frac{1}{n}\sum_{n=1}^{n} \mathbb{I}\left(Y_i \neq f(X_i)\right).$$

The optimal risk, called *Bayes risk*, and the optimal rule are, respectively,

$$R^* = \inf_f R(f), \quad \text{and} \quad f^* = \operatorname*{argmin}_f R(f).$$

Given only finite number of training data $\mathcal{D}_n$, it is natural to consider minimizing the empirical risk over some class of functions $\mathcal{F}$:

$$\widehat{f}_n = \operatorname*{argmin}_{f\in\mathcal{F}} \widehat{R}_n(f).$$

A fundamental question of interest is to know how close is $R(\widehat{f}_n)$ to $R^* = R(f^*)$. The difference, called *excess error*, can be decomposed as:

$$R(\widehat{f}_n) - R^* = \underbrace{\left(R(\widehat{f}_n) - R^*_{\mathcal{F}}\right)}_{\text{estimation error}} + \underbrace{\left(R^*_{\mathcal{F}} - R^*\right)}_{\text{approximation error}},$$

where $R^*_{\mathcal{F}} = \inf_{f\in\mathcal{F}} R(f)$, the best population risk that can be achieved over the class $\mathcal{F}$. The first term is a *random* quantity that reflects the error that we incur because $\widehat{f}_n$ (itself is *random*) is found by minimizing the empirical risk $\widehat{R}_n$ based on n samples, instead of having access to the population risk R. The second term is a *deterministic* quantity that reflects how much we lose by restricting the search space over the class $\mathcal{F}$.

Intuitively, as the sample size n grows, the estimation error should converge to zero (in probability). We are interested in how rapidly it converges, i.e., how the rate depends on the sample size n and the richness of the function class $\mathcal{F}$. Notice that the estimation error can be decomposed further as:

$$R(\widehat{f}_n) - R^*_{\mathcal{F}} = \left(R(\widehat{f}_n) - \widehat{R}_n(\widehat{f}_n)\right) + \left(\widehat{R}_n(\widehat{f}_n) - R^*_{\mathcal{F}}\right).$$

The first term is the difference between empirical risk and population risk of $\widehat{f}_n$, which is trivially bounded by $\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|$. For the second term, we have the following upper bound:

$$\begin{aligned}
\widehat{R}_n(\widehat{f}_n) - R^*_{\mathcal{F}} &= \widehat{R}_n(\widehat{f}_n) - \inf_{f\in\mathcal{F}} R(f) \\
&= \sup_{f\in\mathcal{F}} \left(\widehat{R}_n(\widehat{f}_n) - R(f)\right) \\
&\leq \sup_{f\in\mathcal{F}} \left(\widehat{R}_n(f) - R(f)\right) \\
&\leq \sup_{f\in\mathcal{F}} \left|\widehat{R}_n(f) - R(f)\right|,
\end{aligned}$$

where the first inequality follows from the definition of $\widehat{f}_n$. Overall, the upper bound for the estimation error is:

$$R(\widehat{f}_n) - R^*_{\mathcal{F}} \leq 2 \sup_{f \in \mathcal{F}} \left| \widehat{R}_n(f) - R(f) \right| .$$

For a *fixed* $f \in \mathcal{F}$, $\mathbb{P}(|\widehat{R}_n(f) - R(f)| \geq \epsilon) \to 0$ by the weak law of large numbers. But we'd like to have *uniform convergence* of the empirical risk $\widehat{R}_n(f)$ to the population risk $R(f)$ over a function class $\mathcal{F}$, i.e., we are interested in bounding the probability

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \widehat{R}_n(f) - R(f) \right| \geq \epsilon \right),$$

and it is precisely of the form of uniform bound (1.1) with $Z_i = (X_i, Y_i)$ and

$$g(Z_1, Z_2, \ldots, Z_n) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left(Y_i \neq f(X_i)\right).$$

1.3 Elementary Tail Bounds

Theorem 3 (Markov)**.** *For nonnegative random variable X and $\epsilon > 0$,*

$$\mathbb{P}(X \geq \epsilon) \leq \frac{\mathbb{E}[X]}{\epsilon}. \tag{1.2}$$

Proof. For any nonnegative random variable X, we have inequality $\epsilon \mathbb{I}[X \geq \epsilon] \leq X$. Taking the expectation on both sides and rearranging yields the Markov inequality. □

Theorem 4 (Chebyshev)**.** *For a random variable X with mean μ and finite variance σ^2,*

$$\mathbb{P}(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}. \tag{1.3}$$

Proof. By Markov inequality (1.2),

$$\mathbb{P}(|X - \mu| \geq \epsilon) = \mathbb{P}(|X - \mu|^2 \geq \epsilon^2) \leq \frac{\mathbb{E}[|X - \mu|^2]}{\epsilon^2} = \frac{\sigma^2}{\epsilon^2}.$$

□

Theorem 5 (Chebyshev-Cantelli)**.** *For a random variable X with mean μ and finite variance σ^2, for any $\epsilon > 0$,*

$$\mathbb{P}(X - \mu \geq \epsilon) \leq \frac{\sigma^2}{\sigma^2 + \epsilon^2}, \quad \mathbb{P}(X - \mu \leq -\epsilon) \leq \frac{\sigma^2}{\sigma^2 + \epsilon^2}.$$

It is essentially the one-sided version of Chebyshev's tail bound (1.3).

Proof. For any ϵ, we have

$$(\epsilon - (X - \mu))\,\mathbb{I}[X - \mu < \epsilon] \geq \epsilon - (X - \mu).$$

Taking expectation on both sides yields $\mathbb{E}[(\epsilon - (X - \mu))\,\mathbb{I}[X - \mu < \epsilon]] \geq \epsilon$. For $\epsilon > 0$, applying Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}\epsilon^2 &\leq \mathbb{E}[(\epsilon - (X - \mu))\,\mathbb{I}[X - \mu < \epsilon]]^2 \\ &\leq \mathbb{E}[(\epsilon - (X - \mu))^2]\mathbb{P}(X - \mu < \epsilon) \\ &= \left(\epsilon^2 + \sigma^2\right)\mathbb{P}(X - \mu < \epsilon).\end{aligned}$$

Rearranging the inequality, we have $\mathbb{P}(X - \mu \geq \epsilon) \leq \frac{\sigma^2}{\sigma^2 + \epsilon^2}$. The other side of bound can be obtained by applying the similar procedure to

$$(X - \mu + \epsilon)\,\mathbb{I}[X - \mu > -\epsilon] \geq X - \mu + \epsilon.$$

□

Example 3. Markov's and Chebyshev's inequality cannot be improved in general.

(a) For a given $\epsilon > 0$, let $\mathbb{P}(X = 0) = 1 - p$ and $\mathbb{P}(X = \epsilon) = p$ for any $p \in [0, 1]$. Then

$$\mathbb{P}(X \geq \epsilon) = p = \frac{\mathbb{E}[X]}{\epsilon}.$$

(b) For a given $\epsilon > 0$, let $\mathbb{P}(X = 0) = 1 - p$, $\mathbb{P}(X = \epsilon) = p/2$ and $\mathbb{P}(X = -\epsilon) = p/2$ for some $p \in [0, 1]$. Then $\mu = \mathbb{E}[X] = 0$ and $\sigma^2 = \text{var}[X] = \epsilon^2 p$.

$$\mathbb{P}(|X| \geq \epsilon) = p = \frac{\sigma^2}{\epsilon^2}.$$

Theorem 6 (Chernoff)**.** *For any random variable X and any $s > 0$,*

$$\mathbb{P}(X \geq \epsilon) = \inf_{s>0} \frac{\mathbb{E}[e^{sX}]}{e^{s\epsilon}}. \tag{1.4}$$

Proof. As $s > 0$, we apply Markov inequality (1.2) to the non-negative variable e^{sX},

$$\mathbb{P}(X \geq \epsilon) = \mathbb{P}(e^{sX} \geq e^{s\epsilon}) \leq \frac{\mathbb{E}[e^{sX}]}{e^{s\epsilon}}.$$

Since the above inequality holds for any $s > 0$, taking the infimum with respect to s yields the desired inequality. □

Proposition 7. *The best polynomial moment bound is always at least as tight as the Chernoff bound. Suppose $X \geq 0$ and its moment generating function (mgf) exists in a neighbor of zero. For any $\epsilon > 0$,*

$$\inf_{k=0,1,2,\ldots} \frac{\mathbb{E}[X^k]}{\epsilon^k} \leq \inf_{s>0} \frac{\mathbb{E}[e^{sX}]}{e^{s\epsilon}}. \tag{1.5}$$

Proof. Let $c = \inf_{k=0,1,2,\ldots} \frac{\mathbb{E}[X^k]}{\epsilon^k}$. For any $s > 0$ where its moment generating function exists, we have

$$\begin{aligned} \frac{\mathbb{E}[e^{sX}]}{e^{s\epsilon}} &= \frac{1}{e^{s\epsilon}} \sum_{j=0}^{\infty} \frac{s^j \mathbb{E}[X^j]}{j!} \\ &= \frac{1}{e^{s\epsilon}} \sum_{j=0}^{\infty} \frac{(s\epsilon)^j}{j!} \frac{\mathbb{E}[X^j]}{\epsilon^j} \\ &\geq \frac{c}{e^{s\epsilon}} \sum_{j=0}^{\infty} \frac{(s\epsilon)^j}{j!} \\ &= c. \end{aligned}$$

Taking the infimum over $s > 0$ yields the claim. □

Remark. *Though the best moment bound is shaper than the one obtained by the Chernoff bound, the latter is more widely used and convenient in practice. The exponential form in the mgf $\mathbb{E}[e^{sX}]$ offers more advantages for dealing with sum of independent random variables, as its mgf factorizes into the product of the mgfs of each individual random variable. Accordingly, it is natural to study the behavior (such as its growth rate) of the individual mgf.*

1.4 Sub-Gaussian and Hoeffding Bounds

Definition 3 (Sub-Gaussian)**.** A random variable X with mean $\mu = \mathbb{E}[X]$ is sub-Gaussian with parameter $\tau > 0$ if for all $s \in \mathbb{R}$,

$$\mathbb{E}[\exp(s(X-\mu))] \leq \exp\left\{\frac{\tau^2 s^2}{2}\right\}. \tag{1.6}$$

Remark. *A random variable is sub-Gaussian if and only if its mgf is majorized by the mgf of a Gaussian random variable (see Example 4), hence it is called "sub-Gaussian".*

Proposition 8. *As the mgf exists over the whole line, X has finite moments of all orders: $\mathbb{E}[|X|^k] < \infty$ for any positive integer k. Moreover, $\mathbb{E}[(X-\mu)^2] \leq \tau^2$.*

Proof. As $e^{|X|} \le e^X + e^{-X}$, $\mathbb{E}[e^{|X|}] \le \mathbb{E}[e^X] + \mathbb{E}[e^{-X}] < \infty$. Therefore,

$$\mathbb{E}[|X|^k] \le k!\mathbb{E}[e^{|X|}] < \infty.$$

By Taylor's expansion,

$$\mathbb{E}[\exp(s(X-\mu))] = 1 + \frac{s^2}{2}\mathbb{E}[(X-\mu)^2] + o(s^2).$$

Comparing to

$$\exp\left\{\frac{\tau^2 s^2}{2}\right\} = 1 + \frac{s^2}{2}\tau^2 + o(s^2),$$

and setting $s \to 0$, the inequality (1.6) implies that $\mathbb{E}[(X-\mu)^2] \le \tau^2$. □

Proposition 9. *Any sub-Gaussian variable with parameter τ satisfies the following two-sided exponential tail bound*

$$\mathbb{P}(|X-\mu| \ge \epsilon) \le 2\exp\left\{-\frac{\epsilon^2}{2\tau^2}\right\}. \tag{1.7}$$

Proof. We only prove one-sided bound as the other side is essentially the same. Applying Chernoff bound (1.4) to the random variable $X - \mu$, we obtain

$$\mathbb{P}(X-\mu \ge \epsilon) \le \inf_{s>0} \frac{\mathbb{E}[\exp(s(X-\mu))]}{\exp(s\epsilon)} \le \inf_{s>0} \exp(\tau^2 s^2/2 - s\epsilon) = \exp\left\{-\frac{\epsilon^2}{2\tau^2}\right\}.$$

The last equality is achieved by setting $s = \epsilon/\tau^2 > 0$. □

Proposition 10. *As we will show in Example 4 below, standard Gaussian variable is sub-Gaussian, hence it satisfies the bound (1.7). This bound is sharp up to polynomial-factor corrections, as shown below. Let $Z \sim \mathcal{N}(0,1)$, and $\phi(z) = \frac{1}{\sqrt{2\pi}}\exp(-z^2/2)$ be its density function. For any $z > 0$,*

1. *(Mill's inequality)*

$$\left(\frac{z}{1+z^2}\right)\phi(z) \le \mathbb{P}(Z \ge z) \le \frac{1}{z}\phi(z). \tag{1.8}$$

2. *A variant of the above inequality is*

$$\left(\frac{1}{z} - \frac{1}{z^3}\right)\phi(z) \le \mathbb{P}(Z \ge z) \le \left(\frac{1}{z} - \frac{1}{z^3} + \frac{3}{z^5}\right)\phi(z). \tag{1.9}$$

Proof. See Appendix 1.6.1. □

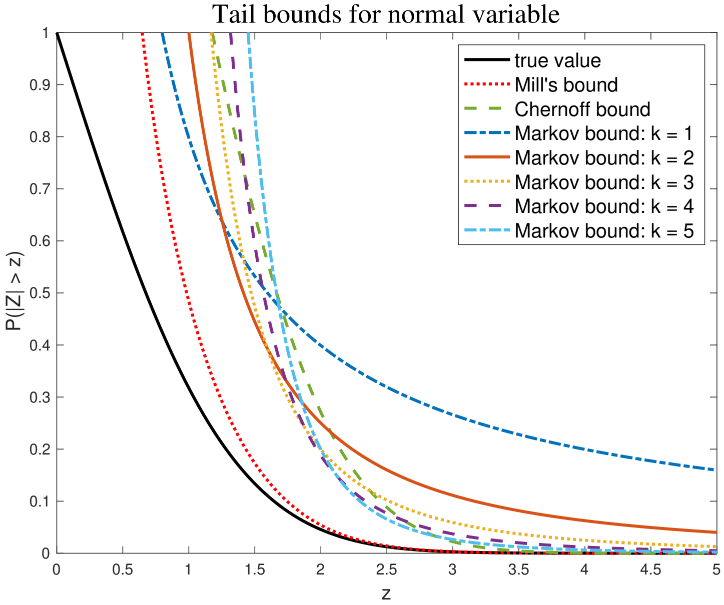

Figure 1.1: A comparison of different upper bounds for $\mathbb{P}(|Z| \geq z)$. Mill's bound: the upper bound in (1.8); Chernoff bound: the bound in (1.7); Markov bound: the bounds in (1.5) for different k.

Remark. *The lower bound in (1.8) is always tighter than the one in (1.9), whereas for large enough z ($z^2 > 3$), the upper bound in (1.9) is tighter than the one in (1.8). See Figure 1.1 for a comparison of different upper bounds for $\mathbb{P}(|Z| \geq z)$. Note that Mill's bound (1.8) is tighter than other generic bounds as it is obtained from the specific form of Gaussian density function (See Appendix 1.6.1).*

Example 4. Classical examples of sub-Gaussian variables are Gaussian, Rademacher, and bounded random variables.

1. (Gaussian) Let $X \sim \mathcal{N}(\mu, \sigma^2)$, its mgf is

$$\mathbb{E}[\exp(sX)] = \exp(\mu s + \sigma^2 s^2/2), \forall s \in \mathbb{R}.$$

Hence (1.6) holds with equality and X is sub-Gaussian with parameter $\tau = \sigma$.

2. (Rademacher) Let X is Rademacher with $\mathbb{P}(X = +1) = \mathbb{P}(X = -1) = 1/2$. Note that $\mathbb{E}[X^k] = 0$ for each odd k and $\mathbb{E}[X^k] = 1$ for each even k. By Taylor expansion,

$$\mathbb{E}[e^{sX}] = \sum_{k=0}^{\infty} \frac{s^k \mathbb{E}[X^k]}{k!} = \sum_{k=0}^{\infty} \frac{s^{2k}}{(2k)!} \leq \sum_{k=0}^{\infty} \frac{s^{2k}}{2^k k!} = e^{s^2/2}, \tag{1.10}$$

where we apply the inequality $2^k k! \leq (2k)!$ in the last step. Hence X is sub-Gaussian with parameter $\tau = 1$.

3. (Bounded) Let X be zero-mean ($\mu = 0$), and supported on $[a, b]$ almost surely. We will use three different ways to show that X is sub-Gaussian, two of which implies its parameter τ is at most $\frac{b-a}{2}$.

   (a) Denote the cumulant generating function of X as $\psi_X(s) = \log \mathbb{E}[e^{sX}]$. It is easy to verify that $\psi_X(0) = 0$ and $\psi'_X(0) = \mu = 0$, and

   $$\psi''_X(s) = \frac{\mathbb{E}[X^2 e^{sX}]}{\mathbb{E}[e^{sX}]} - \left(\frac{\mathbb{E}[X e^{sX}]}{\mathbb{E}[e^{sX}]}\right)^2.$$

   Let Z be a random variable with distribution $dP_s(x) = \frac{e^{sx} dP_X(x)}{\int e^{sx} dP_X(x)}$, then the above identity can be written as

   $$\psi''_X(s) = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2 = \text{var}(Z).$$

   Because P_s (hence Z) is also concentrated on $[a, b]$, its variance is bounded above by:

   $$\psi''_X(s) = \text{var}(Z) = \text{var}\left(Z - \frac{a+b}{2}\right) \leq \frac{(b-a)^2}{4},$$

   where the last inequality follows from the fact that $\left|Z - \frac{a+b}{2}\right| \leq \frac{b-a}{2}$. Putting together, by Taylor's expansion, for some $\xi \in [0, s]$,

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

   which immediately implies that $\mathbb{E}[\exp(sX)] \leq \exp\left\{\frac{s^2(b-a)^2}{8}\right\}$.

   (b) In part (a), we rely on the Taylor's expansion of $\psi_X(s)$ and then bound its second derivative $\psi''_X(s)$. Here, we first apply the convexity of exponential function to obtain the upper bound of $\psi_X(s)$. By identity

   $$sx = \frac{x-a}{b-a} sb + \frac{b-x}{b-a} sa,$$

   we have

   $$e^{sx} \leq \frac{x-a}{b-a} e^{sb} + \frac{b-x}{b-a} e^{sa}, \ \forall x \in [a, b].$$

   Taking expectation on both sides and exploiting $\mathbb{E}[X] = 0$, we obtain

   $$e^{\psi_X(s)} = \mathbb{E}[e^{sX}] \leq \left(1 + \frac{a}{b-a} - \frac{a}{b-a} e^{s(b-a)}\right) e^{sa}.$$

   Denote $p = -a/(b-a)$, the above inequality can be written as

   $$\psi_X(s) \leq \log(1 - p + pe^{s(b-a)}) - ps(b-a).$$

It remains to show $g(s) \doteq \log(1 - p + pe^{s(b-a)}) - ps(b-a) \leq s^2(b-a)^2/8$. By straightforward calculation, $g(0) = g'(0) = 0$, and

$$g''(s) = (b-a)^2 \frac{(1-p)pe^{s(b-a)}}{(1-p+pe^{s(b-a)})^2} \leq \frac{(b-a)^2}{4}.$$

The last inequality follows from the elementary inequality $cd \leq (c+d)^2/4$. Thus, by Taylor's expansion, for some $\xi \in [0, s]$,

$$\psi_X(s) \leq g(s) = g(0) + sg'(0) + \frac{s^2}{2} g''(\xi) \leq \frac{s^2(b-a)^2}{8}.$$

(c) We will apply a technique called *symmetrization* to show that X is sub-Gaussian with parameter at most $\tau = b - a$. First, we introduce an *independent* copy X' of X with $\mathbb{E}_{X'}[X'] = 0$. For any $s \in \mathbb{R}$, we have

$$\mathbb{E}_X[e^{sX}] = \mathbb{E}_X[e^{s(X - \mathbb{E}_{X'}[X'])}].$$

Since the function $f(y) \doteq e^{-sy}$ is convex, by Jensen's inequality,

$$e^{-s\mathbb{E}_{X'}[X']} \leq \mathbb{E}_{X'}[e^{-sX'}].$$

Combining the above two steps, we have

$$\mathbb{E}_X[e^{sX}] \leq \mathbb{E}_{X,X'}[e^{s(X-X')}].$$

Next, we introduce an *independent* Rademacher random variable ϵ. As $X - X'$ is symmetric about 0, the random variabes $(X - X')$ and $\epsilon(X - X')$ have the same distribution. Therefore,

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

Finally, we apply the Rademacher sub-Gaussian bound in (1.10), conditioning on (X, X') to be fixed,

$$\mathbb{E}_X[e^{sX}] \leq E_{X,X',\epsilon}[e^{s\epsilon(X-X')}] = \mathbb{E}_{X,X'}\left[\mathbb{E}_\epsilon[e^{s\epsilon(X-X')}]\right] \leq \mathbb{E}_{X,X'}[e^{\frac{s^2(X-X')^2}{2}}].$$

Because both X and X' are supported on the inverval $[a, b]$, $(X - X')^2 \leq (b-a)^2$. Putting everything together, we conclude

$$\mathbb{E}_X[e^{sX}] \leq e^{\frac{s^2(b-a)^2}{2}},$$

which implies that X is sub-Gaussian with parameter at most $\tau = b - a$.

Theorem 11 (Equivalence of sub-Gaussian variables)**.** *For any zero-mean variable X, all the following properties are equivalent characterizations of being sub-Gaussian:*

(a) Moment generating function: there exists a $\tau > 0$ such that for all $s \in \mathbb{R}$,

$$\mathbb{E}[\exp(sX)] \le \exp\left\{\frac{\tau^2 s^2}{2}\right\}.$$

(b) Majorized by a zero-mean Gaussian random variable: there exists a constant $c \ge 1$ and a Gaussian variable $Z \sim \mathcal{N}(0, \sigma^2)$ such that for all $z \ge 0$,

$$\mathbb{P}(|X| \ge z) \le c\,\mathbb{P}(|Z| \ge z).$$

(c) Moments: there exists a number $\theta \ge 0$ such that for $k = 1, 2, \dots$

$$\mathbb{E}[X^{2k}] \le \frac{(2k)!}{2^k k!}\theta^{2k}.$$

(d) Exponential moments: for all $\lambda \in [0, 1)$,

$$\mathbb{E}\left[\exp\left\{\frac{\lambda X^2}{2\tau^2}\right\}\right] \le \frac{1}{\sqrt{1-\lambda}}.$$

Proof. TODO □

Proposition 12. *Suppose X_1 and X_2 are zero-mean and sub-Gaussian with parameters τ_1 and τ_2 respectively.*

(a) If X_1 and X_2 are independent, $X_1 + X_2$ is sub-Gaussian with parameter $\sqrt{\tau_1^2 + \tau_2^2}$.

(b) In general (without assuming independence), $X_1 + X_2$ is sub-Gaussian with parameter at most $\tau_1 + \tau_2$.

Proof. (a) By independence,

$$\mathbb{E}[\exp(s(X_1 + X_2))] = \mathbb{E}[\exp(sX_1)]\mathbb{E}[\exp(sX_2)] \le \exp\left\{\frac{(\tau_1^2 + \tau_2^2)s^2}{2}\right\}$$

(b) By Hölder inequality, for any $p, q > 1$ such that $p^{-1} + q^{-1} = 1$

$$\begin{aligned}\mathbb{E}[\exp(s(X_1 + X_2))] &\le \left(\mathbb{E}[\exp(psX_1)]\right)^{1/p}\left(\mathbb{E}[\exp(qsX_1)]\right)^{1/q}\\ &\le \exp\left\{\frac{s^2}{2}(p\tau_1^2 + q\tau_2^2)\right\}.\end{aligned}$$

Choosing $p = 1 + \tau_2/\tau_1$ to minimize the bound on the right side, we obtain

$$\mathbb{E}[\exp(s(X_1 + X_2))] \leq \exp\left\{\frac{s^2}{2}(\tau_1 + \tau_2)^2\right\},$$

which establishes the claim.

□

Remark. *If we regard the parameter of a sub-Gaussian variable as its "norm", then (b) implies that it satisfies the desired triangle inequality. See Buldygin and Kozachenko (2000, Theorem 1.2) for more details about the Banach structure of the space of sub-Gaussian random variables.*

Theorem 13 (Hoeffding)**.** *Suppose that $X_i, i = 1, \ldots, n$ are independent sub-Gaussian variables with mean μ_i and parameter τ_i. Then for any $\epsilon > 0$,*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_i)\right| \geq \epsilon\right) \leq 2\exp\left\{-\frac{n\epsilon^2}{\frac{2}{n}\sum_{i=1}^{n}\tau_i^2}\right\}. \tag{1.11}$$

Proof. By Proposition 12(a), $\sum_{i=1}^{n}(X_i - \mu_i)$ is sub-Gaussian with parameter $\sqrt{\sum_{i=1}^{n}\tau_i^2}$. Then the theorem follows Proposition 9 directly. □

Corollary 14. *Let $X_i, i = 1, \ldots, n$ be independent variables with $X_i \in [a_i, b_i]$ almost surely, then for any $\epsilon > 0$,*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])\right| \geq \epsilon\right) \leq 2\exp\left\{-\frac{2n\epsilon^2}{\frac{1}{n}\sum_{i=1}^{n}(b_i - a_i)^2}\right\}.$$

Proof. Setting $\tau_i = \frac{b_i - a_i}{2}$ in (1.11) yields the claim. □

1.5 Sub-exponential

TODO: sub-exponential and Bernstein inequality; Chi-square bound, random projection and Johnson-Lindenstrauss lemma

1.6 Appendix

1.6.1 Proof of Proposition 10

Proof. We will repeatedly apply an elementary identity $\phi'(z) + z\phi(z) = 0$. Hence $\phi'(z) = -z\phi(z) < 0$ for all $z > 0$.

1. For upper bound,

$$\begin{aligned}\mathbb{P}(Z \geq z) &= \int_z^\infty \phi(x)\,dx = \int_z^\infty x\phi(x)\frac{1}{x}\,dx = \int_z^\infty -\phi'(x)\frac{1}{x}\,dx \\ &\leq \int_z^\infty -\phi'(x)\frac{1}{z}\,dx = -\phi(x)|_z^\infty \frac{1}{z} = \frac{1}{z}\phi(z).\end{aligned}$$

For lower bound,

$$\int_z^\infty \phi(x)\,dx \geq \int_z^\infty \phi(x)\left(\frac{x^4+2x^2-1}{x^4+2x^2+1}\right)dx = \frac{-x}{x^2+1}\phi(x)\Big|_z^\infty = \left(\frac{z}{1+z^2}\right)\phi(z).$$

2. Following the proof of upper bound above,

$$\begin{aligned}\mathbb{P}(Z \geq z) &= \int_z^\infty -\phi'(x)\frac{1}{x}\,dx = \int_z^\infty -\frac{1}{x}\,d\,\phi(x) = -\frac{1}{x}\phi(x)\Big|_z^\infty - \int_z^\infty \phi(x)\frac{1}{x^2}\,dx \\ &= \frac{1}{z}\phi(z) - \int_z^\infty x\phi(x)\frac{1}{x^3}\,dx = \frac{1}{z}\phi(z) + \int_z^\infty \phi'(x)\frac{1}{x^3}\,dx,\end{aligned}$$

where we apply $\phi'(z) + z\phi(z) = 0$ twice.
For lower bound, as $\phi'(x) < 0$ for $x > 0$,

$$\mathbb{P}(Z \geq z) \geq \frac{1}{z}\phi(z) + \int_z^\infty \phi'(x)\frac{1}{z^3}\,dx = \frac{1}{z}\phi(z) - \frac{1}{z^3}\phi(z).$$

For upper bound, we apply the integration by part one more time,

$$\begin{aligned}\mathbb{P}(Z \geq z) &= \frac{1}{z}\phi(z) + \int_z^\infty \frac{1}{x^3}\,d\,\phi(x) = \frac{1}{z}\phi(z) + \frac{1}{x^3}\phi(x)\Big|_z^\infty + \int_z^\infty \phi(x)\frac{3}{x^4}\,dx \\ &= \frac{1}{z}\phi(z) - \frac{1}{z^3}\phi(z) - \int_z^\infty \phi'(x)\frac{3}{x^5}\,dx \\ &\leq \frac{1}{z}\phi(z) - \frac{1}{z^3}\phi(z) - \int_z^\infty \phi'(x)\frac{3}{z^5}\,dx \\ &= \frac{1}{z}\phi(z) - \frac{1}{z^3}\phi(z) + \frac{3}{z^5}\phi(z).\end{aligned}$$

□

Bibliography

Buldygin, V. V. and Kozachenko, Y. (2000). *Metric Characterization of Random Variables and Random Processes*. American Mathematical Society. 17

Lehmann, E. L. (2004). *Elements of Large-Sample Theory*. Springer. 7