Lecture 1: Basic Inequalities

Tianpei Xie

Dec. 16th., 2022

Contents

1 Measure Concentraion

- **Remark** The topic of ***measure concentration*** is the study of *random fluctuations of functions of independent random variables.* ***Concentration inequalities*** quantify such statements, typically by bounding the probability that such a function differs from its expected value (or from its median) by more than a certain amount. That is, for small $\epsilon, \delta > 0$

  $$\mathbb{P}\left\{|Z - \mathbb{E}\left[Z\right]| > \epsilon\right\} \le \delta$$

  where $Z = f(X_1, \ldots, X_n)$ is a random variable that smoothly depends on a set of independent random variables $X_1, \ldots, X_n$.

  The main principle, as summarized by Talagrand (1995), is that "*a random variable that **smoothly** depends on the influence of **many independent random variables** satisfies Chernoff type bounds.*"

- **Remark** The concentration-of-measure phenomenon has spread out to an impressively wide range of illustrations and applications, and became a central tool and viewpoint in the quantitative analysis of a number of asymptotic properties in numerous topics of interest including ***geometric analysis***, ***probability theory***, ***statistical mechanics***, ***mathematical statistics*** and ***learning theory***, ***random matrix theory*** or ***quantum information theory***, ***stochastic dynamics***, ***randomized algorithms***, ***complexity***, and so on. [Boucheron et al., 2013]

2 Basic Inequality

2.1 Basic Quantities associated with Random Variables

- Assume a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and a random variable $X : \Omega \to \mathbb{R}$ is a real-valued measurable function on Ω.

- For a random variable X, the ***expectation*** and ***variance*** are denoted as

  $$\mathbb{E}\left[X\right] = \int X d\mathbb{P}$$
  $$Var(X) = \mathbb{E}\left[(X - \mathbb{E}\left[X\right])^2\right]$$

- The ***moment generating function*** of X and its ***logarithm*** are denoted as

  $$M_X(\lambda) := \mathbb{E}\left[e^{\lambda X}\right]$$
  $$\psi_X(\lambda) := \log \mathbb{E}\left[e^{\lambda X}\right]$$

- For $p > 0$, ***the p-th moment** of* X is defined as $\mathbb{E}\left[X^p\right]$, and the ***$p$-th absolute moment*** is $\mathbb{E}\left[|X|^p\right]$.

- The ***L^p norm*** of X is

  $$\|X\|_{L^p} := \mathbb{E}\left[|X|^p\right]^{1/p}$$

where $1 \leq p < \infty$. Note that the L^p space is a *Banach space*, which is defined as

$$L^p(\Omega, \mathbb{P}) := \{X : \|X\|_{L^p} < \infty\}.$$

- *The* ***essential supremum*** of $|X|$ is the L^∞ ***norm*** of X

  $$\|X\|_{L^\infty} := \text{ess sup}\,|X|$$

  Similarly, L^∞ is a Banach space as well

  $$L^\infty(\Omega, \mathbb{P}) := \{X : \|X\|_{L^\infty} < \infty\}.$$

- For $p = 2$, L^2 space is a *Hilbert space* with inner product between random variables $X, Y \in L^2(\Omega, \mathbb{P})$

  $$\langle X\,,\,Y\rangle_{L^2} := \mathbb{E}\left[XY\right] = \int XY d\mathbb{P}$$

  The ***standard deviation*** is

  $$\sigma(X) = (Var(X))^{1/2} = \|X - \mathbb{E}\left[X\right]\|_{L^2}.$$

  The ***covariance*** is defined as

  $$\begin{aligned} cov(X, Y) &:= \langle X - \mathbb{E}\left[X\right]\,,\,Y - \mathbb{E}\left[Y\right]\rangle \\ &= \mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)\left(Y - \mathbb{E}\left[Y\right]\right)\right] \end{aligned}$$

  When we consider random variables as vectors in the Hilbert space L^2, the identity above gives a ***geometric interpretation*** *of the notion of covariance.* The more the vectors $X - \mathbb{E}\left[X\right]$ and $Y - \mathbb{E}\left[Y\right]$ are aligned with each other, the bigger their inner product and covariance are.

- The ***cumulative distribution function (CDF)*** is defined as

  $$F_X(t) := \mathbb{P}\left[X \leq t\right], \quad t \in \mathbb{R}.$$

  The following result is important

  Lemma 2.1 ***(Integral Identity)****. [Vershynin, 2018]*
  Let X be a ***non-negative*** *random variable. Then*

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

  The two sides of this identity are either finite or infinite simultaneously.

2.2 Some Classical Inequalities

- **Proposition 2.2** ***(Jensen's inequality)*** *[Vershynin, 2018]*
  Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space. Let $f : \Omega \to \mathbb{R}$ be a $\mathbb{P}$-measurable function and $\varphi : \mathbb{R} \to \mathbb{R}$ be ***convex function****. Then*

  $$\varphi\left(\mathbb{E}\left[X\right]\right) := \varphi\left(\int X d\mathbb{P}\right) \leq \int \varphi \circ X d\mathbb{P} := \mathbb{E}\left[\varphi\left(X\right)\right]. \tag{2}$$

- **Remark** As a simple consequence of Jensen's inequality, $\|X\|_{L^p}$ is an ***increasing function in*** p, that is

$$\|X\|_{L^p} \le \|X\|_{L^q} \quad \text{for any } 1 \le p \le q \le \infty \tag{3}$$

This inequality follows since $\varphi(x) = x^{q/p}$ is a *convex function* if $q/p \ge 1$.

- **Proposition 2.3** ***(Minkowski's inequality)*** *[Vershynin, 2018]*
For any $p \in [1, \infty]$, $X, Y \in L^p(\Omega, \mathbb{P})$,

$$\|X + Y\|_{L^p} \le \|X\|_{L^p} + \|Y\|_{L^p}, \tag{4}$$

which implies that $\|\cdot\|_{L^p}$ is a norm.

- **Proposition 2.4** ***(Cauchy-Schwarz inequality)*** *[Vershynin, 2018]*
For any random variables $X, Y \in L^2(\Omega, \mathbb{P})$, the following inequality is satisfied:

$$|\langle X\,,\,Y\rangle_{L^2}| := |\mathbb{E}[XY]| \le \|X\|_{L^2}\,\|Y\|_{L^2}. \tag{5}$$

This inequalities can be extended to *conjugate spaces* L^p and L^q

Proposition 2.5 ***(Hölder's inequality)*** *[Vershynin, 2018]*
For $p, q \in (1, \infty)$, $1/p + 1/q = 1$, then the random variables $X \in L^p(\Omega, \mathbb{P})$, $Y \in L^q(\Omega, \mathbb{P})$ satisfy

$$|\langle X\,,\,Y\rangle_{L^2}| := |\mathbb{E}[XY]| \le \|X\|_{L^p}\,\|Y\|_{L^q}. \tag{6}$$

- A classical result is Markov inequality:

Proposition 2.6 ***(Markov's Inequality)***. *[Vershynin, 2018]*
*For any **non-negative** random variable X and $t > 0$, we have*

$$\mathbb{P}\{X \ge t\} \le \frac{\mathbb{E}[X]}{t} \tag{7}$$

Proof: Fix $t > 0$. We can represent any real number x via the identity

$$x = x\mathbb{1}\{x \ge t\} + x\mathbb{1}\{x < t\}$$

Substitute the random variable X for x and *take expectation of both sides.* This gives

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[X\mathbb{1}\{X \ge t\}] + \mathbb{E}[X\mathbb{1}\{X < t\}]\\ &\ge \mathbb{E}[t\,\mathbb{1}\{X \ge t\}] + 0\\ &= t\,\mathbb{P}\{X \ge t\}\end{aligned}$$

Dividing both sides by t, we complete the proof. ∎

- A well-known consequence of *Markov's inequality* is the following *Chebyshev's inequality.* It offers a better, quadratic dependence on t, and instead of the plain tails, it quantifies *the **concentration** of X about its mean.*

Proposition 2.7 ***(Chebyshev's Inequality)***. *[Vershynin, 2018]*
Let X be a random variable with mean μ and variance σ^2. Then, for any $t > 0$, we have

$$\mathbb{P}\{|X - \mu| \ge t\} \le \frac{\sigma^2}{t^2}. \tag{8}$$

- **Remark** If ϕ denotes a ***nondecreasing*** *and* ***nonnegative*** *function* defined on a (possibly infinite) interval $I \subset \mathbb{R}$, and if X denotes a random variable taking values in I, then Markov's inequality implies that for every $t \in I$ with $\phi(t) > 0$,

$$\mathbb{P}\{X \geq t\} \leq \mathbb{P}\{\phi(X) \geq \phi(t)\} \leq \frac{\mathbb{E}[\phi(X)]}{\phi(t)} \tag{9}$$

2.3 Limit Theorems

- **Theorem 2.8** ***(Central Limit Theorem, Linderberg-Leṽy)*** *[Vershynin, 2018]*
Let $X_1, \ldots, X_n$ be ***independent identically distributed*** *random variables with mean $\mathbb{E}[X_i] = 0$ and variance $Var(X_i) = 1$. Then*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \xrightarrow{d} N(0,1) \tag{10}$$

$$i.e. \;\; \lim_{n\to\infty} \sup_{t\in\mathbb{R}} \left| \mathbb{P}\left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \leq t \right\} - \Phi(t) \right| = 0$$

where $\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du = \mathbb{P}\{g \leq t\}$ for some Gaussian variable g.

- **Theorem 2.9** ***(Central Limit Theorem, Nonasymptotic, Berry-Esseen)*** *[Vershynin, 2018]*
Let $X_1, \ldots, X_n$ be ***independent identically distributed*** *random variables with mean $\mathbb{E}[X_i] = 0$, variance $Var(X_i) = \sigma^2$ and $\rho := \mathbb{E}\left[|X_i|^3\right] < \infty$. Then with some constant $C > 0$,*

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}\left\{ \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} X_i \leq t \right\} - \Phi(t) \right| \leq \frac{C}{\sigma^3 \sqrt{n}} \rho \tag{11}$$

where $\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du = \mathbb{P}\{g \leq t\}$ for some Gaussian variable g.

- **Remark** The *Berry-Esseen* version of *central limit theorem* is ***non-asymptotic*** and it has a bound

$$\mathbb{P}\left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \leq t \right\} \leq \mathbb{P}\{g \leq t\} + \frac{C}{\sqrt{n}} \rho = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du + \frac{C}{\sqrt{n}} \rho$$

This bound is ***sharp***, i.e. the equality is attained when $X_i \sim \text{Bernoulli}(1/2)$.

- **Theorem 2.10** ***(Poisson Limit Theorem)****. [Vershynin, 2018]*
Let $X_{N,i}$, $1 \leq i \leq N$, be independent random variables $X_{N,i} \sim Ber(p_{N,i})$, and let $S_N = \sum_{i=1}^{N} X_{N,i}$. Assume that, as $N \to \infty$

$$\max_{i \leq N} p_{N,i} \to 0 \quad \textit{and} \quad \mathbb{E}[S_N] = \sum_{i=1}^{N} p_{N,i} \to \lambda < \infty,$$

Then, as $N \to \infty$,

$$S_N = \sum_{i=1}^{N} X_{N,i} \xrightarrow{d} Pois(\lambda)$$

3 Sum of Independent Random Variables

- The simplest and most thoroughly studied example is *the sum of independent real-valued random variables.* The key to the study of this case is summarized by the trivial but fundamental additive formulas

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}\left(X_i\right)$$

and

$$\psi_{\sum_{i=1}^{n} X_i}(\lambda) = \sum_{i=1}^{n} \psi_{X_i}(\lambda)$$

where $\psi_X(\lambda) := \log \mathbb{E}\left[e^{\lambda X}\right]$ is the logarithm of moment generating function of X.

- **Remark** The consequence of Chebyshevs inequality on the sum of n independent random variables is

$$\mathbb{P}\left\{\frac{1}{n}\left|\sum_{i=1}^{n} \left(X_i - \mathbb{E}\left[X_i\right]\right)\right| \geq t\right\} \leq \frac{\sigma^2}{n\, t^2}$$

where $\sigma^2 := n^{-1} \sum_{i=1}^{n} \mathrm{Var}\left(X_i\right)$.

- **Remark** Choose $\phi(x) = e^{\lambda x}$, we can apply the Markov's inequality to obtain

$$\mathbb{P}\left\{X \geq t\right\} \leq \frac{\mathbb{E}\left[e^{\lambda X}\right]}{e^{\lambda t}} := \frac{M_X(\lambda)}{e^{\lambda t}}. \tag{12}$$

If $Z := X_1 + \ldots + X_n$ as the sum of n independent random variables, then

$$\mathbb{P}\left\{Z - \mathbb{E}\left[Z\right] \geq t\right\} \leq \frac{\mathbb{E}\left[e^{\lambda (Z - \mathbb{E}[Z])}\right]}{e^{\lambda t}} = e^{-\lambda t} \prod_{i=1}^{n} M_{(X_i - \mathbb{E}[X_i])}(\lambda)$$

Note that by union bound

$$\mathbb{P}\left\{\left|Z - \mathbb{E}\left[Z\right]\right| \geq t\right\} \leq \mathbb{P}\left\{Z - \mathbb{E}\left[Z\right] \geq t\right\} + \mathbb{P}\left\{\mathbb{E}\left[Z\right] - Z \geq t\right\}$$

- **Exercise 3.1** *Prove the following inequalities appearing in the text [Boucheron et al., 2013]:*

$$-\log(1-u) - u \leq \frac{u^2}{2(1-u)}, \qquad \textit{for } u \in (0,1), \tag{13}$$

$$h(u) = (1+u)\log(1+u) - u \leq \frac{u^2}{2(1+u/3)}, \qquad \textit{for } u > 0, \tag{14}$$

$$h_1(u) = 1 + u - \sqrt{1+2u} \geq \frac{u^2}{2(1+u)}, \qquad \textit{for } u > 0. \tag{15}$$

3.1 The Cramér-Chernoff Method

- **Remark** In this section we describe and formalize the Cramér-Chernoff bounding method. This method determines *the best possible bound* for a ***tail probability*** that one can possibly obtain using *Markov's inequality* with an exponential function $\phi(t) = e^{\lambda t}$.

  Recall that for a real-valued random variable X, any $\lambda \geq 0$, the following inequality holds

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

  where $\psi_X(\lambda) := \log \mathbb{E}\left[e^{\lambda X}\right]$. One can choose optimal λ^* that ***minimizes*** *the upper bound above*. Since $\psi_X(\lambda)$ is a ***convex function***, we can define its ***Legendre transform***

$$\psi_X^*(t) := \sup_{\lambda \in \mathbb{R}} \left\{\lambda\, t - \psi_X(\lambda)\right\}.$$

  The expression of the right-hand side is known as the ***Fenchel-Legendre dual function*** (or the ***convex conjugate***) of ψ_X. The Legendre transform of log-moment generating function is also its convex conjugate.

- **Proposition 3.2** ***(Chernoff's inequality)*** *[Boucheron et al., 2013]*
  Let X be a real-valued random variable. For $\lambda \geq 0$, $\psi_X(\lambda)$ is the **the logarithm of moment generating function** *of X and $\psi_X^*(t)$ is its* **Legendre (Cramér) transform**. *Then*

$$\mathbb{P}\{X \geq t\} \leq \exp\left(-\psi_X^*(t)\right). \tag{16}$$

- **Remark** The ***Legendre transform*** is also called ***the Cramér transform*** [Boucheron et al., 2013].

  Since $\psi_X(0) = 0$, its *Legendre transform* $\psi_X^*(t)$ is ***nonnegative***.

- **Definition** (***The Rate Function***)
  The rate function is defined as ***the Legendre transformation*** of *the logarithm of the moment generating function* of a random variable. That is,

$$\psi_X^*(t) := \sup_{\lambda \in \mathbb{R}} \left\{\lambda\, t - \psi_X(\lambda)\right\}, \tag{17}$$

  where $\psi_X(\lambda) := \log \mathbb{E}\left[e^{\lambda X}\right]$. Thus, by *Chernoff's inequality*, we can bound *the tail probabilities* of random variables via *its rate function*.

- **Remark** The optimal $\lambda^* := \lambda_t$ that attains the maximum on the right hand side for

$$\psi_X^*(t) = \sup_{\lambda \in \mathbb{R}} \left\{\lambda\, t - \psi_X(\lambda)\right\}$$

  can be found by differentiating $\lambda\, t - \psi_X(\lambda)$ with respect to λ. That is,

$$\psi_X^*(t) = \lambda_t\, t - \psi_X(\lambda_t)$$

  where λ_t is such that $\psi_X'(\lambda_t) = t$. The strict convexity of ψ_X implies that ψ_X' has an ***increasing inverse*** $(\psi_X')^{-1}$ on the interval $\psi_X(I) := (0, B)$ and therefore, for any $t \in (0, B)$,

$$\lambda_t = \left(\psi_X'\right)^{-1}(t).$$

- **Remark** (***Sums of independent random variables***)
The reason why Chernoff's inequality became popular is that it is very simple to use when applied to a sum of independent random variables. As an illustration, assume that $Z := X_1 + \ldots + X_n$ where $X_1, \ldots, X_n$ are ***independent*** *and* ***identically distributed*** *real-valued random variables*. Denote the logarithm of the moment-generating function of the X_i by $\psi_X(\lambda) = \log \mathbb{E}\left[e^{\lambda X_i}\right]$, and the corresponding *Legendre transform* by $\psi_X^*(t)$. Then, by independence, for all λ for which $\psi_X(\lambda) < \infty$,

$$\psi_Z(\lambda) = \log \mathbb{E}\left[e^{\lambda \sum_{i=1}^n X_i}\right] = \log \prod_{i=1}^n \mathbb{E}\left[e^{\lambda X_i}\right] = n\,\psi_X(\lambda)$$

and consequently,

$$\psi_Z^*(t) = n\,\psi_X^*\left(\frac{t}{n}\right).$$

Thus *the Chernoff's inequality* states that

$$\mathbb{P}\{Z \ge t\} \le \exp\left(-\psi_Z^*(t)\right) = \exp\left(-n\,\psi_X^*\left(\frac{t}{n}\right)\right).$$

- **Example** (***Normal Distribution***)
Let X be a ***centered normal random variable*** with variance σ^2. Then

$$\psi_X(\lambda) = \frac{\lambda^2\sigma^2}{2}, \quad \lambda_t = \frac{t}{\sigma^2}$$

and, therefore for every $t > 0$,

$$\psi_X^*(t) = \frac{t^2}{2\sigma^2}.$$

Hence, *Chernoff's inequality* implies, for all $t > 0$,

$$\mathbb{P}\{X \ge t\} \le \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Chernoff's inequality appears to be quite sharp in this case. In fact, one can show that it cannot be improved uniformly by more than a factor of 1/2. ∎

- **Example** (***Poisson Distribution***)
Let X be a ***Poisson random variable*** with parameter ν, that is, $\mathbb{P}\{X = k\} = \frac{1}{k!}e^{\nu}\nu^k$ for all $k = 0, 1, 2, \ldots$ Let $Z = X - \nu$ be the *corresponding centered variable*. Then by direct calculation,

$$\psi_Z(\lambda) = \nu\left(e^{\lambda} - \lambda - 1\right), \quad \lambda_t = \log\left(1 + \frac{t}{\nu}\right)$$

Therefore *the Legendre transform* equals, for every $t > 0$,

$$\psi_Z^*(t) = \nu h\left(\frac{t}{\nu}\right).$$

where the function h is defined, for all $x \ge -1$, by $h(x) = (1+x)\log(1+x) - x$. Similarly, for every $t \le \nu$,

$$\psi_{-Z}^*(t) = \nu h\left(-\frac{t}{\nu}\right).$$

- **Example** (***Bernoulli Distribution***)
  Let X be a ***Bernoulli random variable*** with probability of success p, that is, $\mathbb{P}\{X=1\} = 1-\mathbb{P}\{X=0\} = p$. Let $Z = X - p$ be the *corresponding centered variable.* If $0 < t < 1-p$, we have

$$\psi_Z(\lambda) = \log\left(pe^{\lambda} + 1 - p\right) - p\,\lambda, \qquad \lambda_t = \log\frac{(1-p)(p+t)}{p(1-p-t)}$$

  and therefore, for every $t \in (0, 1-p)$,

$$\psi_Z^*(t) = (1-p-t)\log\frac{1-p-t}{1-p} + (p+t)\log\frac{p+t}{p}.$$

  Equivalently, setting $a = t+p$ for every $a \in (p, 1)$,

$$\psi_Z^*(t) = h_p(a) = (1-a)\log\frac{1-a}{1-p} + a\log\frac{a}{p}.$$

  We note here that $h_p(a)$ is just the ***Kullback-Leibler divergence*** $\mathbb{KL}\left(\mathbb{P}_a \,\|\, \mathbb{P}_p\right)$ between a Bernoulli distribution $\mathbb{P}_a$ of parameter a and a Bernoulli distribution $\mathbb{P}_p$ of parameter p.

$$\mathbb{P}\{X \geq t\} \leq \exp\left(-\mathbb{KL}\left(\mathbb{P}_{p+t} \,\|\, \mathbb{P}_p\right)\right)$$

3.2 Sub-Gaussian Random Variables

- **Definition** (***Sub-Gaussian Random Variable***)
  A ***centered*** random variable X is said to be ***sub-Gaussian** with **variance factor** ν* if

$$\psi_X(\lambda) \leq \frac{\lambda^2\nu}{2}, \qquad \text{for every } \lambda \in \mathbb{R}. \tag{18}$$

  We denote the collection of such random variables by $\mathcal{G}(\nu)$.

- **Remark** Note that this definition does not require the variance of X to be equal to ν, just that it is *bounded by ν*.

  The above definition says that a *centered random variable X* belongs to $\mathcal{G}(\nu)$ if *the moment-generating function* of X is ***dominated by*** *the moment-generating function* of a ***center normal random variable** Y*.

- **Remark** This notion is also convenient because it is naturally *stable under **convolution*** in the sense that if $X_1, \ldots, X_n$ are ***independent*** such that for every i, $X_i \in \mathcal{G}(\nu_i)$, then $\sum_{i=1}^n X_i \in \mathcal{G}(\sum_{i=1}^n \nu_i)$.

- **Remark** (***Characterization***)
  Next we connect the notion of a *sub-Gaussian random variable* with some *other standard ways of defining sub-Gaussian distributions.*

  First observe that *Chernoff's inequality* implies that ***the tail probabilities of a sub-Gaussian random variable are dominated by the corresponding Gaussian tail probabilities***. More precisely, if X belongs to $\mathcal{G}(\nu)$, then for every $t > 0$,

$$\mathbb{P}\{X > t\} \vee \mathbb{P}\{-X > t\} \leq \exp\left(-\frac{t^2}{2\nu}\right)$$

  where $a \vee b$ denotes the *maximum* of a and b.

- **Proposition 3.3** *(**Characterization of Sub-Gaussian Random Variables**) [Boucheron et al., 2013]*
  Let X be a random variable with $\mathbb{E}[X] = 0$. If for some $\nu > 0$

$$\mathbb{P}\{X > t\} \vee \mathbb{P}\{-X > t\} \le \exp\left(-\frac{t^2}{2\nu}\right), \quad \text{for all } t > 0 \tag{19}$$

  then for every integer $q \ge 1$,

$$\mathbb{E}\left[X^{2q}\right] \le 2q!(2\nu)^q \le q!(4\nu)^q. \tag{20}$$

  ***Conversely**, if for some positive constant C*

$$\mathbb{E}\left[X^{2q}\right] \le q!C^q,$$

  then $X \in \mathcal{G}(4C)$ (and therefore (20) holds with $\nu = 4C$).

- **Proposition 3.4** *(**Sub-Gaussian properties**). [Vershynin, 2018]*
  *Let X be a random variable. Then 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.*

  1. *The **tails** of X satisfy*

$$\mathbb{P}\{|X| \ge t\} \le 2\exp\left(-t^2/K_1^2\right) \quad \text{for all } t \ge 0.$$

  2. *The **moments** of X satisfy*

$$\|X\|_{L^p} = \left(\mathbb{E}\left[|X|^p\right]\right)^{1/p} \le K_2\sqrt{p} \quad \text{for all } p \ge 1.$$

  3. *The **moment-generating function (MGF)** of X^2 satisfies*

$$\mathbb{E}\left[\exp(\lambda^2 X^2)\right] \le \exp(K_3^2\,\lambda^2) \quad \text{for all } \lambda \text{ such that } |\lambda| \le \frac{1}{K_3}$$

  4. *The **MGF** of X^2 is **bounded** at some point, namely*

$$\mathbb{E}\left[\exp(X^2/K_4^2)\right] \le 2.$$

  *Moreover, if $\mathbb{E}[X] = 0$ then properties (1)-(4) are also **equivalent** to the following one.*

  5. *The **MGF** of X satisfies*

$$\mathbb{E}\left[\exp(\lambda X)\right] \le \exp(K_5^2\,\lambda^2) \quad \text{for all } \lambda \in \mathbb{R}.$$

- **Remark** (***Equivalent Definitions for Sub-gaussian Random Variables***).
  A random variable X that satisfies one of the equivalent properties (1)-(4) in Proposition above is called a *sub-gaussian random variable.*

  Note that if $\mathbb{E}\left[X^{2q}\right] \le q!C^q$ for every integer q, then setting $\alpha = 1/(2C)$

$$\mathbb{E}\left[\exp(\alpha X^2)\right] = \sum_{q=0}^{\infty} \frac{\alpha^q \mathbb{E}\left[X^{2q}\right]}{q!} \le \sum_{q=0}^{\infty} 2^{-q} = 2$$

Conversely, if

$$\mathbb{E}\left[\exp(\alpha X^2)\right] = \sum_{q=0}^{\infty} \frac{\alpha^q \mathbb{E}\left[X^{2q}\right]}{q!} \le 2$$

then $\sum_{q=1}^{\infty} \frac{\alpha^q \mathbb{E}\left[X^{2q}\right]}{q!} \le 1$, which implies that $\mathbb{E}\left[X^{2q}\right] \le q! \alpha^{-q}$ for every integer q.

- **Definition** (***Sub-Gaussian Norm***)
  The ***sub-gaussian norm*** of X, denoted $\|X\|_{\psi_2}$, is defined to be the ***smallest*** K_4 that satisfies

  $$\mathbb{E}\left[\exp(X^2/K_4^2)\right] \le 2.$$

  In other words, we define

  $$\|X\|_{\psi_2} = \inf\left\{t > 0 : \mathbb{E}\left[\exp(X^2/t^2)\right] \le 2\right\}. \tag{21}$$

- **Remark** (***Sub-Gaussian Properties in Sub-Gaussian Norm***)
  We can restate the properties of sub-gaussian random variables in terms of sub-gaussian norm:

  $$\begin{aligned} \mathbb{P}\left\{|X| \ge t\right\} &\le 2\exp\left(-ct^2/\|X\|_{\psi_2}^2\right) \quad \text{for all } t \ge 0; \\ \|X\|_{L^p} &\le C\|X\|_{\psi_2}\sqrt{p} \quad \text{for all } p \ge 1; \\ \mathbb{E}\left[\exp(X^2/\|X\|_{\psi_2}^2)\right] &\le 2; \\ \text{if } \mathbb{E}[X] = 0, \quad \text{then } \mathbb{E}\left[\exp(\lambda X)\right] &\le \exp(C\lambda^2\|X\|_{\psi_2}^2) \quad \text{for all } \lambda \in \mathbb{R}. \end{aligned}$$

- **Example** Here are some classical examples of sub-gaussian distributions.

  1. (***Gaussian***): As we already noted, $X \sim N(0,1)$ is a sub-gaussian random variable with $\|X\|_{\psi_2} \le C$, where C is an absolute constant. More generally, if $X \sim N(0,\sigma^2)$ then X is sub-gaussian with

     $$\|X\|_{\psi_2} \le C\sigma \tag{22}$$

  2. (***Bernoulli***): Let X be a random variable with ***symmetric Bernoulli distribution***. Since $|X| = 1$, it follows that X is a sub-gaussian random variable with

     $$\|X\|_{\psi_2} \le \frac{1}{\sqrt{\log 2}} \tag{23}$$

  3. (***Bounded***): More generally, any ***bounded random variable*** X is sub-gaussian with

     $$\|X\|_{\psi_2} \le C\|X\|_{\infty} \tag{24}$$

     where $C = 1/\sqrt{\log 2}$.

- **Example** The ***Poisson***, ***exponential***, ***Pareto*** and ***Cauchy*** distributions are ***not sub-gaussian***.

3.3 Sub-Exponential Random Variables

- **Definition** (***Sub-Exponential Random Variables***)
  A ***nonnegative*** random variable X has a ***sub-exponential distribution*** if there exists a constant $a > 0$ such that

  $$\mathbb{E}\left[e^{\lambda X}\right] \le \frac{1}{1-\lambda/a} \quad \text{for every } \lambda \text{ such that } 0 < \lambda < a$$

- **Remark** (***Heavy Tail Distributions***)
  The class of *sub-gaussian distributions* is natural and quite large. Nevertheless, it leaves out some important distributions *whose* ***tails are heavier than gaussian***.

  Consider a standard normal random vector $g = (g_1, \ldots, g_n)$ in $\mathbb{R}_n$, whose coordinates g_i are independent $N(0,1)$ random variables. It is useful in many applications to have a ***concentration inequality*** *for* ***the Euclidean norm*** of g, which is

  $$\|g\|_2 := \left(\sum_{i=1}^{n} g_i^2\right)^{1/2}.$$

  Here we find ourselves in a strange situation. On the one hand, $\|g\|_2$ is a sum of independent random variables g_i^2, so we should expect some concentration to hold. On the other hand, although g_i are *sub-gaussian random variables*, g_i^2 are not. Indeed, recalling the behavior of Gaussian tails we have

  $$\mathbb{P}\left\{g_i^2 > t\right\} = \mathbb{P}\left\{|g_i| > \sqrt{t}\right\} \sim \exp\left(-(\sqrt{t})^2/2\right) = \exp\left(-t^2/2\right)$$

  The tails of g_i^2 are like for the exponential distribution, and are ***strictly heavier than sub-gaussian***.

- **Proposition 3.5** *(**Sub-Exponential properties**). [Vershynin, 2018]*
  Let X be a random variable. Then 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.*

  1. *The* ***tails*** *of X satisfy*

     $$\mathbb{P}\{|X| \ge t\} \le 2\exp(-t/K_1) \quad \textit{for all } t \ge 0.$$

  2. *The* ***moments*** *of X satisfy*

     $$\|X\|_{L^p} = (\mathbb{E}[|X|^p])^{1/p} \le K_2\, p \quad \textit{for all } p \ge 1.$$

  3. *The* ***moment-generating function (MGF)*** *of $|X|$ satisfies*

     $$\mathbb{E}[\exp(\lambda|X|)] \le \exp(K_3\,\lambda) \quad \textit{for all } \lambda \textit{ such that } 0 \le \lambda \le \frac{1}{K_3}$$

  4. *The* ***MGF*** *of $|X|$ is* ***bounded*** *at some point, namely*

     $$\mathbb{E}[\exp(|X|/K_4)] \le 2.$$

  Moreover, if $\mathbb{E}[X] = 0$ then properties (1)-(4) are also ***equivalent*** *to the following one.*

5. *The* ***MGF*** *of X satisfies*

$$\mathbb{E}\left[\exp(\lambda X)\right] \le \exp(K_5^2\,\lambda^2) \quad \text{for all } \lambda \text{ such that } |\lambda| \le \frac{1}{K_5}.$$

- **Remark** (***Equivalent Definitions for Sub-gaussian Random Variables***).
  A random variable X that satisfies one of the equivalent properties (1)-(4) in Proposition above is called a *sub-exponential random variable.*

- **Definition** (***Sub-Exponential Norm***)
  The ***sub-exponential norm*** of X, denoted $\|X\|_{\psi_1}$, is defined to be the ***smallest*** K_4 that satisfies

  $$\mathbb{E}\left[\exp(|X|/K_4)\right] \le 2.$$

  In other words, we define

  $$\|X\|_{\psi_1} = \inf\left\{t > 0 : \mathbb{E}\left[\exp(|X|/t)\right] \le 2\right\}. \tag{25}$$

- **Remark** Sub-gaussian and sub-exponential distributions are closely related.

  1. First, *any sub-gaussian distribution is clearly sub-exponential.*
  2. Second, *the* ***square*** *of a* ***sub-gaussian random variable*** *is* ***sub-exponential***:

     Lemma 3.6 *(**Sub-exponential is Sub-gaussian Squared**). [Vershynin, 2018]*
     A random variable X is **sub-gaussian** *if and only if X^2 is* **sub-exponential**. *Moreover,*

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

     More generally, *the* ***product*** *of two* ***sub-gaussian random variables*** *is* ***sub-exponential***:

     Lemma 3.7 *(**Product of Sub-Gaussians is Sub-Exponential**). [Vershynin, 2018]*
     Let X and Y be ***sub-gaussian random variables****. Then XY is* ***sub-exponential****. Moreover,*

     $$\|XY\|_{\psi_1} \le \|X\|_{\psi_2}\,\|Y\|_{\psi_2}.$$

- **Proposition 3.8** *(**Characterization of Sub-Exponential Random Variables**) [Boucheron et al., 2013]*
  Let X be a nonnegative random variable. If X is sub-exponential distributed with parameter $a > 0$ then for every integer $q \ge 1$,

  $$\mathbb{E}\left[X^q\right] \le 2^{q+1}\frac{q!}{a^q}. \tag{26}$$

  Conversely*, if there exists a constant $a > 0$ in order that for every positive integer q,*

  $$\mathbb{E}\left[X^q\right] \le \frac{q!}{a^q},$$

  then X is sub-exponential. More precisely, for any $0 < \lambda < a$,

  $$\mathbb{E}\left[e^{\lambda X}\right] \le \frac{1}{1 - \lambda/a}.$$

- **Example** Here are some classical examples of sub-exponential distributions.

  1. (***Exponential***): Recall that X has ***exponential distribution*** with rate $a > 0$, denoted $X \sim Exp(a)$, if X is a *non-negative random variable* with tails

  $$\mathbb{P}\{X \geq t\} \leq \exp(-at), \quad \forall t \geq 0$$

  Then

  $$\|X\|_{\psi_1} \leq \frac{C}{a} \tag{27}$$

3.4 Sub-Gamma Random Variables

- **Remark** For *exponential distribution* $X \sim \exp(a)$ with *rate* a (*inverse* of *scale parameter*), the p.d.f. and moment generating function

  $$f_X(x) = ae^{-ax}, \quad x > 0$$
  $$M_X(\lambda) = \frac{1}{1-\lambda/a}, \quad 0 < \lambda < a$$

  For *Gamma distribution* $X \sim \Gamma(a, 1/b)$ with *shape parameter* a and *scale parameter* b, the p.d.f. and the moment generating function

  $$f_X(x) = \frac{1}{\Gamma(a)\, b^a} x^{a-1} e^{-x/b}, \quad x > 0$$
  $$M_X(\lambda) = \left(\frac{1}{1-b\lambda}\right)^a, \quad 0 < \lambda < 1/b$$

  Also $\mathbb{E}[X] = ab$ and $\mathrm{Var}(X) = ab^2$.

- **Definition** (***Sub-Gamma Random Variables***)
  A real-valued centered random variable X is said to be <u>***sub-gamma** on **the right tail***</u> *with **variance factor** ν and **scale parameter** c* if

  $$\psi_X(\lambda) \leq \frac{\lambda^2 \nu}{2(1-c\lambda)} \quad \text{for every } \lambda \text{ such that } 0 < \lambda < 1/c$$

  We denote the collection of such random variables by $\Gamma_+(\nu, c)$.

  Similarlly, a real-valued centered random variable X is said to be <u>***sub-gamma** on **the left tail***</u> *with **variance factor** ν and **scale parameter** c* if $-X$ is ***sub-gamma on the right tail*** *with variance factor ν and tail parameter c.* We denote the collection of such random variables by $\Gamma_-(\nu, c)$.

  Finally, X is simply said to be <u>***sub-gamma***</u> *with **variance factor** ν and **scale parameter** c* if X is *sub-gamma **both** on the right and left tails* with ***the same*** *variance factor* ν and *scale parameter* c. The collection of such random variables is denoted by $\Gamma(\nu, c)$.

  Observe that $\Gamma(\nu, 0) = \mathcal{G}(\nu)$.

- **Remark** To derive the definition fo sub-gamma distribution, we see that *the variance factor* $\nu := ab^2$ and $c := b$. Also $\mathbb{E}[X] = ab$. The logarithmic moment generating function of Gamma distribution $\Gamma(a, 1/b) = \Gamma(\nu/c^2, 1/c)$ is

$$\psi_{X-\mathbb{E}[X]}(\lambda) = a \log\left(\frac{1}{1-b\lambda}\right) - \lambda ab \le \frac{\lambda^2 b^2 a}{2(1-b\lambda)} \equiv \frac{\lambda^2 \nu}{2(1-c\lambda)}$$

The last inequality is due to

$$\log\left(\frac{1}{1-u}\right) - u \le \frac{u^2}{2(1-u)}$$

- **Remark** Note that the sum of n i.i.d. random variables with exponential distribution $\exp(1/b)$ have the Gamma distribution $\Gamma(n, 1/b)$. So ***the sub-gamma distributed*** random variable follows ***the sub-exponential distribution*** as well (with shape parameter = 1).

- **Remark** (***Characterization***)
Similarly to the *sub-Gaussian property*, the ***sub-gamma property*** can be characterized in terms of *tail or moment conditions.* We start by computing ***the Fenchel-Legendre dual function*** of

$$\psi(\lambda) = \frac{\lambda^2 \nu}{2(1-c\lambda)}.$$

Setting

$$h_1(u) = 1 + u - \sqrt{1+2u} \text{ for } u > 0,$$

it follows by elementary calculation that for every $t > 0$,

$$\psi^*(t) = \sup_{\lambda \in (0, 1/c)} \left\{ t\lambda - \frac{\lambda^2 \nu}{2(1-c\lambda)} \right\} = \frac{\nu}{c^2} h_1\left(\frac{c\,t}{\nu}\right).$$

Since h_1 is an increasing function from $(0, \infty)$ onto $(0, \infty)$ with ***inverse function***

$$h^{-1}(u) = u + \sqrt{2u} \text{ for } u > 0,$$

we finally get

$$\psi^{*-1}(u) = \sqrt{2\nu u} + cu.$$

Hence, *Chernoff's inequality* implies that whenever X is a *sub-gamma random variable on the right tail* with *variance factor* ν and *scale parameter* c, for every $t > 0$, we have

$$\mathbb{P}\{X > t\} \le \exp\left(\frac{\nu}{c^2} h_1\left(\frac{c\,t}{\nu}\right)\right), \tag{28}$$

or equivalently, for every $t > 0$,

$$\mathbb{P}\left\{X > \sqrt{2\nu t} + ct\right\} \le e^{-t}. \tag{29}$$

Therefore, if X belongs to $\Gamma(\nu, c)$, then for every $t > 0$,

$$\mathbb{P}\left\{X > \sqrt{2\nu t} + ct\right\} \vee \mathbb{P}\left\{-X > \sqrt{2\nu t} + ct\right\} \le e^{-t}. \quad \blacksquare$$

3.5 Orlicz Spaces

- **Definition** (***Orlicz Spaces***) [Vershynin, 2018]
A function : $\psi : [0, \infty) \to [0, \infty)$ is called ***an Orlicz function*** if ψ is ***convex***, ***increasing***, and satisfies:

$$\psi(0) = 0, \quad \psi(x) \to \infty, \text{ as } x \to \infty.$$

For a *given Orlicz function* ψ, ***the Orlicz norm*** *of a random variable* X is defined as

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

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

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

- **Example** The Orlicz Spaces generalizes the L^p space for random variables:

  1. L^p ***space***: Consider the function

  $$\psi(x) = x^p,$$

  which is obviously *an Orlicz function* for $p \geq 1$. The resulting Orlicz space L_{ψ} is the classical space L^p.

  2. L_{ψ} ***space***: Consider the function

  $$\psi_2(x) = e^{x^2} - 1,$$

  which is obviously *an Orlicz function*. The resulting *Orlicz norm* is exactly the sub-gaussian norm $\|\cdot\|_{\psi_2}$ that we defined. The corresponding Orlicz space L_{ψ_2} consists of *all sub-gaussian random variables*.

- **Remark** We can easily locate L_{ψ_2} in the hierarchy of the classical L^p spaces:

$$L^{\infty} \subset L_{\psi_2} \subset L^p \text{ for every } p \in [1, \infty).$$

Thus *the space of* ***sub-gaussian random variables*** L_{ψ_2} is smaller than all of L^p spaces, but it is still larger than *the space of* ***bounded random variables*** L^{∞}.

3.6 A Maximal Inequality

- The purpose of this section is to show how information about the Legendre transform of random variables in a finite collection can be used to bound the expected maximum of these random variables.

- **Remark** (***Mean of Maximum of Finite Sub-Gaussian Random Variables***)
Let $X_1, \ldots, X_n$ be real-valued random variables where a $\nu > 0$ exists such that for every $i = 1, \ldots, n$, the logarithm of the moment-generating function of X_i satisfies

$$\psi_{X_i}(\lambda) \leq \frac{\lambda^2 \nu}{2}, \quad \text{for all } \lambda > 0.$$

Then, by Jensen's inequality,

$$
\begin{aligned}
\exp\left(\lambda \mathbb{E}\left[\max_{i=1,\dots,n} X_i\right]\right) &\le \mathbb{E}\left[\exp\left(\lambda \max_{i=1,\dots,n} X_i\right)\right] \\
&= \mathbb{E}\left[\max_{i=1,\dots,n} \exp\left(\lambda X_i\right)\right] \\
&\le \sum_{i=1}^{n} \mathbb{E}\left[\exp\left(\lambda X_i\right)\right] \\
&\le n \exp\left(\frac{\lambda^2 \nu}{2}\right)
\end{aligned}
$$

Taking logarithms on both sides, we have

$$
\mathbb{E}\left[\max_{i=1,\dots,n} X_i\right] \le \frac{\log n}{\lambda} + \frac{\lambda \nu}{2}
$$

The upper bound is minimized for $\lambda^* = \sqrt{2 \log n / \nu}$, which yields

$$
\mathbb{E}\left[\max_{i=1,\dots,n} X_i\right] \le \sqrt{2\nu \log n} \tag{30}
$$

This simple bound is ***asymptotically sharp*** if the X_i are ***i.i.d. normal random variables***

- **Lemma 3.9** ***(Generalized Inverse of Legendre Transform)*** *[Boucheron et al., 2013]*
  Let ϕ be a ***convex*** *and* ***continuously differentiable*** *function defined on the interval $[0, b)$ where $0 < b \le \infty$. Assume that $\phi(0) = \phi'(0) = 0$ and set, for every $t \ge 0$, the Legendre transform*

  $$
  \phi^*(t) = \sup_{\lambda \in (0,b)} \left\{ \lambda\, t - \phi(\lambda) \right\}.
  $$

  Then ϕ^* is a ***nonnegative convex*** *and* ***nondecreasing function*** *on $[0, \infty)$. Moreover, for every $y \ge 0$, the set $\{t \ge 0 : \phi^*(t) > y\}$ is* ***non-empty*** *and* <u>***the generalized inverse of ϕ^****</u>*, defined by*

  $$
  \phi^{*-1}(y) = \inf \left\{ t \ge 0 : \phi^*(t) > y \right\}, \tag{31}
  $$

  can also be written as

  $$
  \phi^{*-1}(y) = \inf_{\lambda \in (0,b)} \left[\frac{y + \phi(\lambda)}{\lambda} \right].
  $$

- The next result offers a convenient bound for the expected value of the maximum of finitely many exponentially integrable random variables. This type of bound has been used in so-called ***chaining arguments*** *for bounding* ***suprema of Gaussian or empirical processes****.*

  Proposition 3.10 ***(Mean of Maximum of Finite Random Variables with Convex Bounds on Log-MGF)*** *[Boucheron et al., 2013]*
  Let $X_1, \dots, X_n$ be real-valued random variables such that for every $\lambda \in (0, b)$ and $i = 1, ..., n$, the ***logarithm of the moment-generating function*** *of X_i satisfies*

  $$
  \psi_{X_i}(\lambda) \le \phi(\lambda) \tag{32}
  $$

*where ϕ is a **convex** and **continuously differentiable** function on $[0, b)$ with $0 < b \le \infty$ such that $\phi(0) = \phi'(0) = 0$. Then*

$$\mathbb{E}\left[\max_{i=1,\dots,n} X_i\right] \le \phi^{*-1}(\log n). \tag{33}$$

*In particular, if the X_i are **sub-Gaussian with variance factor** ν, that is, $\psi_{X_i}(\lambda) \le \lambda^2\nu/2$ for every $\lambda \in (0, \infty)$, then*

$$\mathbb{E}\left[\max_{i=1,\dots,n} X_i\right] \le \sqrt{2\nu \log n}.$$

Proof: By Jensen's inequality,

$$\exp\left(\lambda \mathbb{E}\left[\max_{i=1,\dots,n} X_i\right]\right) \le \mathbb{E}\left[\exp\left(\lambda \max_{i=1,\dots,n} X_i\right)\right] = \mathbb{E}\left[\max_{i=1,\dots,n} \exp\left(\lambda X_i\right)\right]$$

for any $\lambda \in (0, b)$. Thus, recalling that $\psi_{X_i}(\lambda) = \log \mathbb{E}\left[\exp(\lambda X_i)\right]$,

$$\exp\left(\lambda \mathbb{E}\left[\max_{i=1,\dots,n} X_i\right]\right) \le \sum_{i=1}^{n} \mathbb{E}\left[\exp\left(\lambda X_i\right)\right] \le n \exp\left(\phi(\lambda)\right).$$

Therefore, for any $\lambda \in (0, b)$,

$$\lambda \mathbb{E}\left[\max_{i=1,\dots,n} X_i\right] - \phi(\lambda) \le \log n,$$

which means that

$$\mathbb{E}\left[\max_{i=1,\dots,n} X_i\right] \le \inf_{\lambda \in (0,b)} \left\{\frac{\log n + \phi(\lambda)}{\lambda}\right\}$$

and the results follows from the equivalent definition of generalized inverse of ϕ. ∎

- **Corollary 3.11** ***(Mean of Maximum of Finite Sub-Gamma Random Variables)***
  Let $X_1, \dots, X_n$ be real-valued random variables belonging to $\Gamma_+(\nu, c)$. Then

$$\mathbb{E}\left[\max_{i=1,\dots,n} X_i\right] \le \sqrt{2\nu \log n} + c \log n. \tag{34}$$

3.7 Hoeffding's Inequality

- **Remark** (***Bounded Variables***)
  Bounded variables are an important class of *sub-Gaussian random variables*. The *sub-Gaussian property* of *bounded random variables* is established by the following lemma:

- **Lemma 3.12** ***(Hoeffding's Lemma)*** *[Boucheron et al., 2013]*
  *Let X be a random variable with $\mathbb{E}[X] = 0$, taking values in a **bounded interval** $[a, b]$ and let $\psi_X(\lambda) := \log \mathbb{E}\left[e^{\lambda X}\right]$. Then*

$$\psi''_X(\lambda) \le \frac{(b-a)^2}{4}$$

and $X \in \mathcal{G}((b-a)^2/4)$.

Proof: Observe first that

$$\left| X - \frac{b+a}{2} \right| \leq \frac{b-a}{2}$$

and therefore

$$\mathrm{Var}(X) = \mathrm{Var}\left(X - \frac{(b+a)}{2} \right) \leq \frac{(b-a)^2}{4}.$$

Now, let $\mathbb{P}$ denote the distribution of X and let $\mathbb{P}_\lambda$ be the probability distribution with density

$$x \mapsto e^{-\psi_X(\lambda)} e^{\lambda x}$$

with respect to $\mathbb{P}$. Since $\mathbb{P}_\lambda$ is *concentrated* on $[a, b]$, the variance of a random variable Z with distribution $\mathbb{P}_\lambda$ is *bounded by* $(b-a)^2/4$. Hence, by an elementary computation,

$$\begin{aligned} \psi''_X(\lambda) &= e^{-\psi_X(\lambda)} \mathbb{E}\left[X^2 e^{\lambda X} \right] - e^{-2\psi_X(\lambda)} \left(\mathbb{E}\left[X e^{\lambda X} \right] \right)^2 \\ &= \mathrm{Var}(Z) \leq \frac{(b-a)^2}{4}. \end{aligned}$$

The sub-Gaussian property follows by noting that $\psi_X(0) = \psi'_X(0) = 0$, and by *Taylor's theorem* that implies that, for some $\theta \in [0, \lambda]$,

$$\psi_X(\lambda) = \psi_X(0) + \lambda\, \psi'_X(0) + \frac{\lambda^2}{2} \psi''_X(\theta) \leq \frac{\lambda^2 (b-a)^2}{8}. \quad \blacksquare$$

- **Proposition 3.13** ***(Hoeffding's inequality)*** *[Boucheron et al., 2013]*
  Let $X_1, \ldots, X_n$ be independent random variables such that X_i takes its values in $[a_i, b_i]$ ***almost surely*** *for all $i \leq n$. Let*

  $$S = \sum_{i=1}^{n} \left(X_i - \mathbb{E}\left[X_i \right] \right).$$

  Then for every $t > 0$,

  $$\mathbb{P}\left\{ S \geq t \right\} \leq \exp\left(- \frac{2t^2}{\sum_{i=1}^{n} (b_i - a_i)^2} \right). \tag{35}$$

- **Remark** (***Hoeffding's inequality for Scaled Radmatcher Random Variables***)
  Let us consider the random variables

  $$X_i = \epsilon_i \alpha_i, \quad i = 1, \ldots, n$$

  where $\epsilon_1, \ldots, \epsilon_n$ are ***independent Rademacher random variables*** (i.e. ***symmetric Bernoulli random variables*** with $\mathbb{P}\left\{ \epsilon_i = 1 \right\} = \mathbb{P}\left\{ \epsilon_i = -1 \right\} = 1/2$) and $\alpha_1, \ldots, \alpha_n$ are real numbers. We get

  $$\mathbb{P}\left\{ S \geq t \right\} \leq \exp\left(- \frac{t^2}{2 \sum_{i=1}^{n} \alpha_i^2} \right).$$

- **Remark** (***Hoeffding's inequality*** as ***Concentration version of the Central Limit Theorem***) [Vershynin, 2018]
  We can view Hoeffding's inequality as *a* ***concentration version*** *of* ***the central limit theorem***. Indeed, the most we may expect from a concentration inequality is that ***the tail of*** $\sum_i \epsilon_i \alpha_i$ behaves similarly to ***the tail of the normal distribution***.

  With the normalization $\|\alpha\|_2 = 1$, Hoeffdings inequality provides the tail $e^{-t^2/2}$, which is exactly the same as the bound for the standard normal tail. This is good news. We have been able to obtain the same exponentially light tails for sums as for the normal distribution, even though *the difference of these two distributions is not exponentially small.*

- **Remark** (***Non-asymptotic Results***). [Vershynin, 2018]
  It should be stressed that unlike ***the classical limit theorems*** of ***Probability Theory***, *Hoeffding's inequality* is ***non-asymptotic*** in that it holds for ***all fixed*** N *as opposed to* $N \to \infty$. The ***larger*** N, the ***stronger*** inequality becomes. As we will see later, ***the non-asymptotic nature*** of ***concentration inequalities*** like *Hoeffding* makes them attractive in application in data sciences, where N often corresponds to ***sample size***.

- **Proposition 3.14** ***(General Hoeffding's inequality)*** *[Vershynin, 2018]*
  Let $X_1, \ldots, X_n$ be ***independent sub-gaussian*** *random variables. Let*

$$S = \sum_{i=1}^{n} \left(X_i - \mathbb{E}\left[X_i\right]\right).$$

  Then for every $t > 0$,

$$\mathbb{P}\left\{S \geq t\right\} \leq \exp\left(-\frac{c\, t^2}{\sum_{i=1}^{n} \|X_i\|_{\psi_2}}\right). \tag{36}$$

  Proposition 3.15 ***(General Hoeffding's inequality, Linear Form)*** *[Vershynin, 2018]*
  Let $X_1, \ldots, X_n$ be ***independent, mean zero, sub-gaussian*** *random variables, and $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$. Then for every $t > 0$,*

$$\mathbb{P}\left\{\sum_{i=1}^{n} a_i X_i \geq t\right\} \leq \exp\left(-\frac{c\, t^2}{K^2\, \|a\|_2^2}\right). \tag{37}$$

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

- **Proposition 3.16** ***(Khintchine's inequality, L^p norm,*** *$p \geq 2$). [Vershynin, 2018]*
  Let $X_1, \ldots, X_n$ be ***independent sub-gaussian*** *random variables with* ***zero means*** *and* ***unit variances****, and let $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$. Then for every $p \in [2, \infty)$, we have*

$$\left(\sum_{i=1}^{n} a_i^2\right)^{1/2} \leq \left\|\sum_{i=1}^{n} a_i X_i\right\|_{L^p} \leq C\, K \sqrt{p} \left(\sum_{i=1}^{n} a_i^2\right)^{1/2} \tag{38}$$

  where $K = \max_i \|X_i\|_{\psi_2}$ and C is an absolute constant.

  Proposition 3.17 ***(Khintchine's inequality, L^1 norm)****. [Vershynin, 2018]*
  Let $X_1, \ldots, X_n$ be ***independent sub-gaussian*** *random variables with* ***zero means*** *and* ***unit***

variances*, and let* $a = (a_1, \dots, a_n) \in \mathbb{R}^n$*. Then*

$$c(K)\left(\sum_{i=1}^{n} a_i^2\right)^{1/2} \le \left\|\sum_{i=1}^{n} a_i X_i\right\|_{L^1} \le \left(\sum_{i=1}^{n} a_i^2\right)^{1/2} \tag{39}$$

where $K = \max_i \|X_i\|_{\psi_2}$ *and* $c(K) > 0$ *is a quantity which may depend only on* K*.*

3.8 Bennett's Inequality

- **Remark** Our starting point is the fact that *the logarithmic moment-generating function* of an *independent sum* equals *the sum of the logarithmic moment-generating functions of the centered summands*, that is,

$$\psi_S(\lambda) = \sum_{i=1}^{n}\left(\log \mathbb{E}\left[e^{\lambda X_i}\right] - \lambda \mathbb{E}\left[X_i\right]\right).$$

  Using $\log u \le u - 1$ for $u > 0$,

$$\psi_S(\lambda) \le \sum_{i=1}^{n} \mathbb{E}\left[e^{\lambda X_i} - \lambda X_i - 1\right]. \tag{40}$$

  Both Bennett's and Bernstein's inequalities may be derived from this bound, under different integrability conditions for the X_i.

- **Proposition 3.18** ***(Bennett's Inequality)*** *[Boucheron et al., 2013]*
  Let $X_1, \dots, X_n$ *be independent random variables with* ***finite variance*** *such that* $X_i \le b$ *for some* $b > 0$ ***almost surely*** *for all* $i \le n$*. Let*

$$S = \sum_{i=1}^{n}\left(X_i - \mathbb{E}\left[X_i\right]\right)$$

  and $\nu = \sum_{i=1}^{n} \mathbb{E}\left[X_i^2\right]$*. If we write* $\phi(u) = e^u - u - 1$ *for* $u \in \mathbb{R}$*, then, for all* $\lambda > 0$*,*

$$\log \mathbb{E}\left[e^{\lambda S}\right] \le n \log\left(1 + \frac{\nu}{nb^2}\phi(b\lambda)\right) \le \frac{\nu}{b^2}\,\phi(b\lambda),$$

  and for any $t > 0$*,*

$$\mathbb{P}\left\{S \ge t\right\} \le \exp\left(-\frac{\nu}{b^2} h\left(\frac{b\,t}{\nu}\right)\right) \tag{41}$$

  where $h(u) = (1 + u)\log(1 + u) - u$ *for* $u > 0$*.*

- **Remark** This bound can be analyzed in two different regimes:

  1. *In the* ***small deviation regime***, where $u := bt/\nu \ll 1$, we have *asymptotically* $h(u) \approx u^2$ and *Bennett's inequality* gives *approximately the Gaussian tail bound* $\approx \exp(-t^2/\nu)$.

  2. *In the* ***large deviations regime***, say where $u := bt/\nu \ge 2$, we have $h(u) \ge \frac{1}{2} u \log u$, and *Bennett's inequality* gives a ***Poisson-like tail*** $(\nu/bt)^{t/2b}$.

3.9 Bernstein's Inequality

- **Definition** (***Bernstein's Condition***)
  Given a *random variable* X with mean $\mu = \mathbb{E}[X]$ we say that ***Bernstein's condition*** *with parameter* ν, c holds if the *variance* $\mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - \mu^2 \le \nu$, and
  $$\sum_{i=1}^{n} \mathbb{E}\left[(X-\mu)_+^q\right] \le \frac{q!}{2}\nu c^{q-2}, \quad \text{for all integers } q \ge 2,$$
  where $(x)_+ = \max\{x, 0\}$.

- **Proposition 3.19** *(**Bernstein's Condition $\Rightarrow$ Sub-Gamma Distribution**). [Boucheron et al., 2013]*
  Let $X_1, \ldots, X_n$ be independent real-valued random variables. Assume that there exist positive numbers ν and c such that $\sum_{i=1}^{n} \mathbb{E}\left[X_i^2\right] \le \nu$ and
  $$\sum_{i=1}^{n} \mathbb{E}\left[(X_i)_+^q\right] \le \frac{q!}{2}\nu c^{q-2}, \quad \textit{for all integers } q \ge 3,$$
  where $(x)_+ = \max\{x, 0\}$. If $S = \sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])$, then for all $\lambda \in (0, 1/c)$ and $t > 0$
  $$\psi_S(\lambda) \le \frac{\lambda^2 \nu}{2(1 - c\lambda)}$$
  and
  $$\psi_S^*(t) \ge \frac{\nu}{c^2} h_1\left(\frac{ct}{\nu}\right),$$
  where $h_1(u) = 1 + u - \sqrt{1+2u}$ for $u > 0$. In particular, for all $t > 0$,
  $$\mathbb{P}\left\{S \ge \sqrt{2\nu t} + ct\right\} \le e^{-t}. \tag{42}$$

  Remark Note that
  $$h_1(u) = 1 + u - \sqrt{1+2u} = \frac{1}{2}\left(\sqrt{1+2u} - 1\right)^2;$$
  $$h_1^{-1}(s) = \frac{1}{2}(\sqrt{2s} + 1)^2 - 1 = s + \sqrt{2s}$$

- **Corollary 3.20** *(**Bernstein's Inequality, Sub-Gamma Distribution**). [Boucheron et al., 2013]*
  Let $X_1, \ldots, X_n$ be independent real-valued random variables satisfying the conditions above and let $S = \sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])$. Then for all $t > 0$,
  $$\mathbb{P}\{S \ge t\} \le \exp\left(-\frac{t^2}{2(\nu + ct)}\right). \tag{43}$$

- **Remark** For bounded random variables $X_i \le b$ almost surely for all $i \le n$, then the conditions of above corollary hold with
  $$\nu = \sum_{i=1}^{n} \mathbb{E}\left[X_i^2\right], \quad c = b/3.$$

- We are ready to state and prove a concentration inequality for sums of independent sub-exponential random variables.

  Proposition 3.21 ***(Bernstein's Inequality)****. [Vershynin, 2018]*
  Let $X_1, \ldots, X_n$ be ***independent, mean zero, sub-exponential random variables****. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left\{\left|\sum_{i=1}^{n} X_i\right| \geq t\right\} \leq 2\exp\left[-c\min\left\{\frac{t^2}{\sum_{i=1}^{n}\|X_i\|_{\psi_2}^2}, \frac{t}{\max_i \|X_i\|_{\psi_1}}\right\}\right] \tag{44}$$

  where $c > 0$ is an absolute constant.

- **Proposition 3.22** ***(Bernstein's Inequality, Linear Combination Form)****. [Vershynin, 2018]*
  Let $X_1, \ldots, X_n$ be ***independent, mean zero, sub-exponential random variables****, and $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left\{\left|\sum_{i=1}^{n} a_i X_i\right| \geq t\right\} \leq 2\exp\left[-c\min\left\{\frac{t^2}{K^2\,\|a\|_2^2}, \frac{t}{K\,\|a\|_\infty}\right\}\right] \tag{45}$$

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

- **Corollary 3.23** ***(Bernstein's Inequality, Average Form)****. [Vershynin, 2018]*
  Let $X_1, \ldots, X_n$ be ***independent, mean zero, sub-exponential random variables****. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \geq t\right\} \leq 2\exp\left[-c\min\left\{\frac{t^2}{K^2}, \frac{t}{K}\right\} n\right] \tag{46}$$

  where $K = \max_i \|X_i\|_{\psi_1}$.

- **Remark** This bound can be analyzed in two different regimes:

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \geq t\right\} \leq \begin{cases} 2\exp\left(-ct^2\right) & t \leq C\sqrt{n} \\ 2\exp\left(-t\sqrt{n}\right) & t \geq C\sqrt{n} \end{cases}$$

  1. *In the* ***small deviation regime***, where $t \leq C\sqrt{n}$, we have a ***sub-gaussian tail bound*** as if the sum had a normal distribution with constant variance. Note that *this domain* ***widens*** as n ***increases*** and ***the central limit theorem*** becomes more powerful.

  2. *In the* ***large deviations regime***, say where $t \geq C\sqrt{n}$, the sum has a *heavier, sub-exponential tail bound*, which can be due to the contribution of ***a single term*** X_i.

- **Proposition 3.24** ***(Bernstein's Inequality for Bounded Distributions)****. [Vershynin, 2018]*
  Let $X_1, \ldots, X_n$ be ***independent, mean zero*** *random variables, such that $|X_i| \leq b$ all i. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \geq t\right\} \leq 2\exp\left(-\frac{t^2}{2(\nu + bt/3)}\right). \tag{47}$$

  Here $\nu = \sum_{i=1}^{n}\mathbb{E}\left[X_i^2\right]$ is the variance of the sum.

3.10 The Johnson-Lindenstrauss Lemma

- **Remark** (***Overview of The Johnson-Lindenstrauss Lemma***)
  The celebrated ***Johnson-Lindenstrauss lemma*** states roughly that, given an arbitrary set of n *points* in a (high-dimensional) Euclidean space, there exists a ***linear embedding*** of these points in a *d-dimensional Euclidean space* such that ***all pairwise distances are preserved*** within a factor of $1 \pm \epsilon$ if d is proportional to $(\log n)/\epsilon^2$. It is remarkable that *this result* ***does not involve the dimension of the space*** to which the n points belong. In fact, the dimension of this space may even be *infinite*.

- **Definition** (***ϵ-Isometry***)
  Consider an arbitrary set $A \subset \mathbb{R}^D$ or $A \subset \mathcal{H}$ for *separable Hilbert space $\mathcal{H}$*. Given $\epsilon \in (0, 1)$, a map $f : \mathbb{R}^D \to \mathbb{R}^d$ is called an ***ϵ-isometry on A*** if for every pair $a, a' \in A$, we have

$$(1-\epsilon)\left\|a-a'\right\|_2^2 \le \left\|f(a)-f(a')\right\|_2^2 \le (1+\epsilon)\left\|a-a'\right\|_2^2.$$

- **Remark** (***Problem Statement***)
  A natural question is to find ***the smallest possible value of*** d for which *a linear ϵ-isometry exists on A. The Johnson-Lindenstrauss lemma*, stated and proved below, ensures that when A is a ***finite set*** with cardinality n, a *linear ϵ-isometry* exists whenever $d \ge \kappa\epsilon^{-2}\log(n)$, where κ is an absolute constant.

- **Remark** (***Gaussian Random Projection***)
  The basic idea is to construct a ***random projection*** $W : \mathbb{R}^D \to \mathbb{R}^d$ (i.e. a *linear mapping*) that is an exact ***isometry*** "***in expectation***," that is, for every $\alpha \in \mathbb{R}^D$,

$$\mathbb{E}\left[\|W(\alpha)\|_2^2\right] = \mathbb{E}\left[\|\alpha\|_2^2\right].$$

  In other words, denoting by $L^{2,d}$ the space of *square-integrable $\mathbb{R}^d$-valued random vectors*, W is an ***isometry*** from $\mathbb{R}^D$ into $L^{2,d}$.

  To construct W, let $X_{i,j}$, $i = 1, \ldots, d$, $j = 1, \ldots, D$ be *independent and identically distributed* real-valued random variables such that $\mathbb{E}[X_{i,j}] = 0$ and $\mathrm{Var}(X_{i,j}) = 1$. For every $\alpha = (\alpha_1, \ldots, \alpha_D) \in \mathbb{R}^D$ and $i \in \{1, \ldots, d\}$, define

$$W_i(\alpha) := \sum_{j=1}^{D} \alpha_j X_{i,j}$$

  $W_i(\alpha)/\sqrt{d}$ is the i-th component of the random vector $W(\alpha)$, that is, W is defined by

$$W(\alpha) := \left(\frac{1}{\sqrt{d}}\sum_{j=1}^{D}\alpha_j X_{i,j}\right)_{i=1}^{d}$$
$$\Rightarrow W(\alpha) := \frac{1}{\sqrt{d}} X\alpha^T.$$

  Observe that by independence of the $X_{i,j}$, for every $i = 1, \ldots, d$,

$$\mathbb{E}\left[W_i(\alpha)^2\right] = \mathbb{E}\left[\left(\sum_{j=1}^{D}\alpha_j X_{i,j}\right)^2\right] = \sum_{j=1}^{D}\alpha_i^2\,\mathbb{E}\left[X_{i,j}^2\right] = \mathbb{E}\left[\|\alpha\|_2^2\right].$$

Therefore, for every $\alpha \in \mathbb{R}^D$,

$$\mathbb{E}\left[\|W(\alpha)\|_2^2\right] = \frac{1}{d}\sum_{i=1}^{d} \mathbb{E}\left[W_i(\alpha)^2\right] = \mathbb{E}\left[\|\alpha\|_2^2\right].$$

and indeed, W is an *isometry* from $\mathbb{R}^D$ into $L^{2,d}$.

- **Theorem 3.25** ***(The Johnson-Lindenstrauss Lemma)*** *[Boucheron et al., 2013]*
  Let A be a ***finite subset*** *of $\mathbb{R}^D$ with cardinality n. Assume that for some $\nu \geq 1$, $X_{i,j} \in \mathcal{G}(\nu)$ and let $\epsilon, \delta \in (0,1)$. If*

$$d \geq 100 \frac{\nu^2 \log\left(\frac{n}{\sqrt{\delta}}\right)}{\epsilon^2},$$

then with probability at least $1-\delta$, W is an ***ϵ-isometry*** *on A. That is, for every pair $a, a' \in A$, with probability $1-\delta$ we have*

$$(1-\epsilon)\left\|a - a'\right\|_2^2 \leq \left\|\frac{1}{\sqrt{d}} X a^T - \frac{1}{\sqrt{d}} X (a')^T\right\|_2^2 \leq (1+\epsilon)\left\|a - a'\right\|_2^2.$$

Proof: Denote by $\mathbb{S}$ the unit sphere of $\mathbb{R}^D$ and let T be the subset of $\mathbb{S}$ defined by

$$T := \left\{\frac{a - a'}{\|a - a'\|_2} : a, a' \in A, a \neq a'\right\}$$

Then T has cardinality $N \leq n(n-1)/2$. We need to show that, under the stated condition for d,

$$\sup_{\alpha \in T}\left|\|W(\alpha)\|^2 - 1\right| \leq \epsilon.$$

First note that for all $\alpha \in \mathbb{S}$ and $i \leq d$, using the fact that the $X_{i,j}$ are *sub-Gaussian*,

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda W_i(\alpha)\right)\right] &= \mathbb{E}\left[\exp\left(\lambda \sum_{j=1}^{D} \alpha_j X_{i,j}\right)\right] \\
&= \prod_{j=1}^{D} \mathbb{E}\left[\exp\left(\lambda \alpha_j X_{i,j}\right)\right] \\
&\leq \mathbb{E}\left[\exp\left(\frac{\nu\lambda^2 \sum_{j=1}^{D} \alpha_j^2}{2}\right)\right] \\
&= \mathbb{E}\left[\exp\left(\frac{\nu\lambda^2}{2}\right)\right], \qquad (\text{since } \alpha \in \mathbb{S},\ \|\alpha\|^2 = 1)
\end{aligned}$$

and therefore $W_i(\alpha) \in \mathcal{G}(\nu)$. Thus, by moment characterization of sub-Gaussian random variable, for every integer $q \geq 2$,

$$\mathbb{E}\left[W_i(\alpha)^{2q}\right] \leq \frac{q!}{2} \times 4(2\nu)^q \leq \frac{1}{2} q! (4\nu)^q.$$

Hence, since for each α the random variables $W_i(\alpha)$, $i = 1, \ldots, d$ are *independent*, we may use *Bernstein's inequality* for $\sum_{i=1}^d W_i(\alpha)^2$ with $d(4\nu) \to \nu$ and $d\nu \to c$ to obtain, for every $\alpha \in T$ and $t > 0$,

$$\mathbb{P}\left\{\left|\sum_{i=1}^{d}\left(W_i(\alpha)^2 - 1\right)\right| \geq 4\nu\sqrt{2dt} + d\nu t\right\} \leq 2e^{-t}$$

This implies, by the union bound,

$$\mathbb{P}\left\{\sup_{\alpha \in T}\left|\sum_{i=1}^{d}\left(W_i(\alpha)^2 - 1\right)\right| \geq 4\nu\left(\sqrt{2dt} + dt\right)\right\} \leq 2Ne^{-t} \leq n^2 e^{-t}$$

Setting $t = \log(n^2/\delta) = 2\log(n/\sqrt{\delta})$,

$$4\nu\left(\sqrt{2dt} + dt\right) = 8\nu\left(\sqrt{d\log\left(n/\sqrt{\delta}\right)} + d\log\left(n/\sqrt{\delta}\right)\right) \geq 8\nu\sqrt{d\log\left(n/\sqrt{\delta}\right)}$$

we have

$$\mathbb{P}\left\{\sup_{\alpha \in T}\left|\sum_{i=1}^{d}\left(W_i(\alpha)^2 - 1\right)\right| \geq 8\nu\sqrt{d\log\left(n/\sqrt{\delta}\right)}\right\} \leq \delta$$

or, equivalently,

$$\mathbb{P}\left\{\sup_{\alpha \in T}\left|\|W(\alpha)\|^2 - 1\right| \geq \sqrt{\frac{8\nu\log\left(n/\sqrt{\delta}\right)}{d}} + \frac{8\nu\log\left(n/\sqrt{\delta}\right)}{d}\right\} \leq \delta$$

Finally, we see that $d \geq 100\frac{\nu^2\log\left(\frac{n}{\sqrt{\delta}}\right)}{\epsilon^2}$ implies that

$$\sqrt{\frac{8\nu\log\left(n/\sqrt{\delta}\right)}{d}} + \frac{8\nu\log\left(n/\sqrt{\delta}\right)}{d} \leq \frac{4\epsilon}{5} + \frac{2\epsilon^2}{25\nu} \leq \epsilon$$

and therefore, with probability at least $1 - \delta$:

$$\sup_{\alpha \in T}\left|\|W(\alpha)\|^2 - 1\right| \leq \epsilon \quad \blacksquare$$

References

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.