

# *High-Dimensional Probability: Answers, Theorems, and Definitions*

Last revised on August 11, 2021

- Companion notes for *High-Dimensional Probability*, by Roman Vershynin. Link to book (PDF available online): [www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-book.html](http://www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-book.html).
- **Disclaimer:** These notes compile my answers to the exercises, and lift the required theorems and definitions from the book. I wrote these notes to aid my personal study of the book. Read them at your own risk!\*

## Contents

<b>0</b>	<b>Appetizer: Using probability to cover a geometric set</b>	<b>2</b>
<b>1</b>	<b>Preliminaries on random variables</b>	<b>5</b>
1.1	Basic quantities . . . . .	5
1.2	Some classical inequalities . . . . .	5
1.3	Limits theorems . . . . .	7
<b>2</b>	<b>Concentrations of sums of independent random variables</b>	<b>8</b>
2.1	Why concentration inequalities? . . . . .	8
2.2	Hoeffding's inequality . . . . .	9
2.3	Chernoffs's inequality . . . . .	11

---

\*Scribe: Alex Bie, [alexbie98@gmail.com](mailto:alexbie98@gmail.com).

## 0 Appetizer: Using probability to cover a geometric set

A point  $x \in \mathbb{R}^n$  is a **convex combination** of points  $x_1, \dots, x_m \in \mathbb{R}^n$  if

$$x = \sum_{i=1}^m \lambda_i x_i \quad \text{with each } \lambda_i \geq 0 \quad \text{and} \quad \sum_{i=1}^m \lambda_i = 1.$$

The **convex hull** of  $T \subseteq \mathbb{R}^n$ ,  $\text{conv}(T)$ , is the set of all convex combinations of  $T$ .

**Theorem 0.0.1** (Carathéodory's Theorem). Let  $x \in \text{conv}(T)$ . There exists  $k \leq n + 1$  points  $x_1, \dots, x_k \in T$  such that  $x$  is a convex combination of  $x_1, \dots, x_k$ .

The result says we can obtain any point in the convex hull of  $T$  using at most a dimension-dependent number of points. Let the **diameter** of a set  $T$  be defined as  $\text{diam}(T) = \sup\{\|x - y\|_2 : x, y \in T\}$ .

**Theorem 0.0.2** (Approximate Carathéodory's Theorem). Let  $\text{diam}(T) = 1$ . Let  $x \in \text{conv}(T)$ . For any  $k$ , there exists  $k$  points  $x_1, \dots, x_k \in T$  such that

$$\left\| x - \frac{1}{k} \sum_{j=1}^k x_j \right\|_2 \leq \frac{1}{\sqrt{k}}$$

*Proof.* Suppose  $|T| = m$ . WLOG we can assume  $T$  is bounded by 1 in  $\|\cdot\|_2$ . We write  $x = \sum_{i=1}^m \lambda_i x_i$ , and interpret  $\lambda_i$  as probabilities. We define the random variable

$$X = x_i \text{ with probability } \lambda_i$$

for  $i = 1, \dots, m$ . We can verify that  $\mathbb{E}X = \sum_{i=1}^m \lambda_i x_i = x$ . Taking  $X_1, \dots, X_k \stackrel{\text{iid}}{\sim} X$ . It remains to analyse the quantity  $\mathbb{E}\|x - \frac{1}{k} \sum_{j=1}^k X_j\|_2^2$ .

$$\begin{aligned} \mathbb{E} \left\| x - \frac{1}{k} \sum_{j=1}^k X_j \right\|_2^2 &\leq \frac{1}{k^2} \mathbb{E} \left\| \sum_{j=1}^k X_j - x \right\|_2^2 \\ &= \frac{1}{k^2} \sum_{j=1}^k \mathbb{E} \|X_j - x\|_2^2 && \text{(by Exercise 0.0.3 (a))} \\ &= \frac{1}{k} \mathbb{E} \|X - x\|_2^2 \end{aligned}$$

Applying the result of Exercise 0.0.3 (b), we obtain

$$\mathbb{E}\|X - x\|_2^2 = \mathbb{E}\|X\|_2^2 - \|\mathbb{E}X\|_2^2 \leq \mathbb{E}\|X\|_2^2 \leq 1$$

Plugging this in above, we obtain the desired bound in expectation, hence there must exist a realization of the  $X_j$ ,  $x_1, \dots, x_k$ , such that the bound holds.  $\square$

**Exercise 0.0.3.** Check the following identities for random vectors.

(a) Let  $X_1, \dots, X_k$  be independent, mean zero random vectors in  $\mathbb{R}^n$ . Show that

$$\mathbb{E} \left\| \sum_{j=1}^k X_j \right\|_2^2 = \mathbb{E} \sum_{j=1}^k \|X_j\|_2^2$$

*Answer.*

$$\begin{aligned}
\mathbb{E} \left\| \sum_{j=1}^k X_j \right\|_2^2 &= \sum_{i=1}^n \mathbb{E} \left( \sum_{j=1}^m X_j^{(i)} \right)^2 = \sum_{i=1}^n \text{Var} \left( \sum_{j=1}^m X_j^{(i)} \right) && \text{(by mean zero)} \\
&= \sum_{i=1}^n \sum_{j=1}^m \text{Var} \left( X_j^{(i)} \right) && \text{(by independence)} \\
&= \sum_{i=1}^n \sum_{j=1}^m \mathbb{E} \left( X_j^{(i)} \right)^2 && \text{(by mean zero)} \\
&= \mathbb{E} \sum_{j=1}^m \|X_j\|_2^2
\end{aligned}$$

□

Among other things, this result implies that the expected squared distance of a random walk (starting from the origin) is equal to sum of the expected squared distances of each step.

(b) Let  $X$  be a random vector in  $\mathbb{R}^n$ . Show that

$$\mathbb{E} \|X - \mathbb{E}X\|_2^2 = \mathbb{E} \|X\|_2^2 - \|\mathbb{E}X\|_2^2$$

*Answer.*

$$\begin{aligned}
\mathbb{E} \|X - \mathbb{E}X\|_2^2 &= \mathbb{E} \sum_{i=1}^n \left( X^{(i)} - (\mathbb{E}X)^{(i)} \right)^2 = \sum_{i=1}^n \text{Var}(X^{(i)}) = \sum_{i=1}^n \mathbb{E} \left( X^{(i)} \right)^2 - \left( \mathbb{E}X^{(i)} \right)^2 \\
&= \mathbb{E} \|X\|_2^2 - \|\mathbb{E}X\|_2^2
\end{aligned}$$

□

**Corollary 0.0.4** (Covering polytopes by balls). Let  $P \subseteq \mathbb{R}^n$  be a polytope with  $\text{diam}(P) = 1$ . Let  $m$  be the number of vertices of  $P$ . Let  $\varepsilon > 0$ . We can cover  $P$  with  $m^k$  balls of radius  $\varepsilon$  for  $k \geq \lceil 1/\varepsilon^2 \rceil$ .

*Proof.* Take  $T$  to be the vertex set of  $P$ .  $|T| = m$ . Note that for any  $x \in P$ ,  $x \in \text{conv}(T)$ . By Theorem 0.0.2, taking  $k \geq \lceil 1/\varepsilon^2 \rceil$ , we can find  $x_1, \dots, x_k \in T$  such that

$$\left\| x - \frac{1}{k} \sum_{j=1}^k x_j \right\| \leq \frac{1}{\sqrt{k}} \leq \varepsilon$$

The number of ball centres obtained from selecting a set of  $k$  points out of  $m$  with repetition is bounded by  $m^k$  (possibly repeating orders). Hence we have an  $\varepsilon$ -cover sufficient to cover  $P$ . □

**Exercise 0.0.5** (Binomial coefficient inequality). Show that for  $1 \leq r \leq n$

$$\left( \frac{n}{r} \right)^r \leq \binom{n}{r} \leq \sum_{k=0}^r \binom{n}{k} \leq \left( \frac{en}{r} \right)^r$$

*Answer.* For the first inequality, consider

$$\frac{\left( \frac{n}{r} \right)^r}{\binom{n}{r}} = \frac{\frac{n}{r} \cdot \frac{n}{r} \cdot \dots \cdot \frac{n}{r}}{\frac{n}{r} \cdot \frac{n-1}{r-1} \cdot \dots \cdot \frac{n-r+1}{1}} \leq 1 \cdot 1 \cdot \dots \cdot 1 = 1$$

The second inequality follows immediately. To justify the last inequality, write

$$\begin{aligned}
\left(\frac{en}{r}\right)^r &= e^r \cdot \left(\frac{n}{r}\right)^r = \sum_{k=0}^{\infty} \frac{r^k}{k!} \cdot \left(\frac{n}{r}\right)^r && \text{(Maclaurin series for } e^x\text{)} \\
&\geq \sum_{k=0}^r \frac{r^k}{k!} \cdot \left(\frac{n}{r}\right)^r \\
&= \sum_{k=0}^r \frac{n^k \cdot n^{r-k}}{k! \cdot r^{r-k}} \\
&\geq \sum_{k=0}^r \frac{n^k}{k!} && \text{(by } n \geq r\text{)} \\
&\geq \sum_{k=0}^r \binom{n}{k}
\end{aligned}$$

□

**Exercise 0.0.6** (Improved covering). Show that in the setting of Corollary 0.0.4, for  $k \geq \lceil 1/\varepsilon^2 \rceil$

$$(C + C\varepsilon^2 m)^k$$

balls suffice for a suitable constant  $C$ .

*Answer.* We can give a tighter bound than given in the proof of Corollary 0.0.4 on the number of ball centres obtained from selecting a set of  $k$  points out of  $m$  with repetition (since computing the mean of  $k$  is order-invariant with respect to input points). By the “stars-and-bars”<sup>†</sup> argument, this quantity is given by

$$\binom{m+k-1}{k-1}$$

Note that  $\min\{k-1, m\} = k-1 \leq \min\{k, m-1\}$ , so looking at row  $m+k-1$  of Pascal’s triangle

$$\binom{m+k-1}{k-1} \leq \binom{m+k-1}{k}$$

Then, using Exercise 0.0.5

$$\binom{m+k-1}{k} \leq \left(\frac{e(m+k-1)}{k}\right)^k = \left(e\frac{k-1}{k} + e\frac{1}{k}m\right)^k \leq (e + e\varepsilon^2 m)^k$$

□

---

<sup>†</sup>[https://en.wikipedia.org/wiki/Stars\\_and\\_bars\\_\(combinatorics\)](https://en.wikipedia.org/wiki/Stars_and_bars_(combinatorics))

# 1 Preliminaries on random variables

## 1.1 Basic quantities

The **expectation** of a random variable  $X$  is denoted as  $\mathbb{E}X$ , and **variance** is denoted as  $\text{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2$ . (We note that the expectation operator  $\mathbb{E}$  can be directly defined as the Lebesgue integral of the random variable (measurable function)  $X : \Omega \rightarrow \mathbb{R}$  in the probability space  $(\Omega, \mathcal{M}, \mathbb{P})$ .)

The **moment generating function** of  $X$  is given by

$$M_X(t) = \mathbb{E}e^{tX} \quad \text{for all } t \in \mathbb{R}$$

The **p-th moment** of  $X$  is given by  $\mathbb{E}X^p$ . We also let  $\|X\|_p = (\mathbb{E}X^p)^{\frac{1}{p}}$  denote the **p-norm** of  $X$ . For  $p = \infty$ , we have

$$\|X\|_\infty = \text{ess sup } X$$

recalling that the **essential supremum** of a function  $f$  is the "smallest value  $\gamma$  such that  $\{\omega \in \Omega : |f(\omega)| > \gamma\}$  has measure 0".

From this, we can define the  **$L^p$  spaces**<sup>†</sup>, given a probability space  $(\Omega, \mathcal{M}, \mathbb{P})$

$$L^p = \{X : \|X\|_p < \infty\}$$

Results from measure and integration theory tell us that the  $(L^p, \|\cdot\|_p)$  are complete. In the case of  $L^2$ , we have that with the inner product

$$\begin{aligned} \langle X, Y \rangle &= \int_{\Omega} XY(\omega) \mu(\omega) \\ &= \mathbb{E}XY \end{aligned}$$

$(L^2, \langle \cdot, \cdot \rangle)$  is a Hilbert space. In this case we can express the **standard deviation** of  $X$  as  $\sqrt{\text{Var}(X)} = \|X - \mathbb{E}X\|_2$ , and the **covariance** of random variable  $X$  and  $Y$  as

$$\text{Cov}(X, Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y) = \langle X - \mathbb{E}X, Y - \mathbb{E}Y \rangle$$

In this setting, considering random variables as vectors in  $L^2$ , the covariance between  $X$  and  $Y$  can be interpreted as the *alignment* between the vectors  $X - \mathbb{E}X$  and  $Y - \mathbb{E}Y$ .

## 1.2 Some classical inequalities

We say  $f : \mathbb{R} \rightarrow \mathbb{R}$  is **convex** if

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y) \quad \text{for all } x, y \in \mathbb{R} \text{ and } t \in [0, 1]$$

**Jensen's inequality** states that for any random variable  $X$  and a convex function  $f$ , we get

$$f(\mathbb{E}X) \leq \mathbb{E}(f(X))$$

A corollary of Jensen's inequality implies that<sup>§</sup>

$$\|X\|_p \leq \|X\|_q \quad \text{for all } 1 \leq p \leq q \leq \infty$$

**Minkowski's inequality** asserts that the triangle inequality holds for the  $L_p$  spaces

$$\|X + Y\|_p \leq \|X\|_p + \|Y\|_p \quad \text{for all } X, Y \in L^p$$

In  $L^2$ , we have the **Cauchy-Schwarz inequality**, which states that  $|\mathbb{E}XY| \leq \mathbb{E}|XY| \leq \|X\|_2\|Y\|_2$ . **Holder's inequality** additionally asserts that for  $1/p + 1/q = 1$

$$|\mathbb{E}XY| \leq \|XY\|_1 \leq \|X\|_p\|Y\|_q$$

<sup>†</sup>A technical note is that the objects of  $L_p$  are actually equivalence classes of functions  $[X]$  with equality almost everywhere, otherwise  $\|\cdot\|_p$  is only a semi-norm.

<sup>§</sup>For  $q < \infty$ , the result follows by applying Jensen's inequality for  $f(x) = x^{\frac{q}{p}}$ . Otherwise,  $\|X\|_\infty = \gamma = (\mathbb{E}\gamma^p)^{\frac{1}{p}} = \|\gamma\|_p \geq \|X\|_p$ .

which also holds for  $p = 1, q = \infty$ .

The **cumulative distribution function** of  $X$  is defined as

$$F_X(t) = \mathbb{P}\{X \leq t\} = \mathbb{P}(X^{-1}(-\infty, t]) \quad \text{for all } t \in \mathbb{R}$$

and we refer to  $\mathbb{P}\{X > t\} = 1 - F_X(t)$  as the **tail** of  $X$ .

**Lemma 1.2.1** (Integral identity). Let  $X \geq 0$  be a random variable. Then

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

with left side  $= \infty$  iff right side  $= \infty$ .

**Exercise 1.2.2** (Generalization of integral identity). Show that Lemma can be extended to be valid for any  $X$

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

*Answer.* For not necessary non-negative  $X$ ,  $\mathbb{E}X := \mathbb{E}X^+ - \mathbb{E}X^-$  when they exist and are both  $< \infty$ , where

$$X^+ = \begin{cases} X & \text{if } X \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad X^- = \begin{cases} -X & \text{if } X \leq 0 \\ 0 & \text{otherwise} \end{cases}$$

Applying Lemma 1.2.1 to the terms yields the result. For the second term

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

□

**Exercise 1.2.3** ( $p$ -th moment via the tail). Let  $X$  be a random variable and  $0 < p < \infty$ . Show that

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

whenever the right side is  $< \infty$ .

*Answer.* On the right side, substitute  $u = t^p$ , so  $du = pt^{p-1} dt$  and

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

where the last equality comes from applying Lemma 1.2.1 to the random variable  $|X|^p \geq 0$ .

□

**Proposition 1.2.4** (Markov's inequality). Let  $X \geq 0$  with  $\mathbb{E}X < \infty$ . Then for  $t > 0$

$$\mathbb{P}\{X \geq t\} \leq \frac{\mathbb{E}X}{t}$$

*Proof.* Fix  $t > 0$ . Applying Lemma 1.2.1

$$\mathbb{E}X = \int_0^\infty \mathbb{P}\{X \geq x\} dx \geq \int_0^t \mathbb{P}\{X \geq x\} dx \geq \int_0^t \mathbb{P}\{X \geq t\} dx = t \cdot \mathbb{P}\{X \geq t\}$$

□

**Corollary 1.2.5** (Chebyshev's inequality). Let  $X$  have  $\mathbb{E}X < \infty$  and  $\text{Var}(X) < \infty$ . Then for  $t > 0$

$$\mathbb{P}\{|X - \mathbb{E}X| > t\} \leq \frac{\text{Var}(X)}{t^2}$$

**Exercise 1.2.6.** Give a proof of Chebyshev's inequality using Markov's inequality.

*Answer.* The random variable  $|X - \mathbb{E}X|^2$  is well-defined (by  $\mathbb{E}X < \infty$ ), non-negative, with finite expectation. Applying Markov's inequality with  $t^2 > 0$  yields

$$\mathbb{P}\{|X - \mathbb{E}X| \geq t\} = \mathbb{P}\{|X - \mathbb{E}X|^2 \geq t^2\} \leq \frac{\text{Var}(X)}{t^2}$$

□

### 1.3 Limits theorems

For independent and identically distributed variables  $X_1, \dots, X_N$ , the sample mean  $\frac{1}{N} \sum_{i=1}^N X_i$  has

$$\text{Var}\left(\frac{1}{N} \sum_{i=1}^N X_i\right) = \frac{\text{Var}(X_1)}{N} \rightarrow 0 \text{ as } N \rightarrow \infty$$

so we should expect it to concentrate around the true mean.

**Theorem 1.3.1** (Strong law of large numbers). Let  $X_1, X_2, \dots$  be a sequence of i.i.d. random variables with  $\mathbb{E}X_1 < \infty$ . Then the averaged partial sums

$$\frac{S_N}{N} = \frac{1}{N} \sum_{i=1}^N X_i \rightarrow \mathbb{E}X_1 \text{ almost surely}$$

where random variables  $(Y_N)_{N=1}^\infty$  are said to **converge almost surely** to a random variable  $Y$  if there exists measurable  $Z \in \mathcal{M}$  with  $\mathbb{P}(Z) = 0$  and

$$\lim_{N \rightarrow \infty} Y_N(\omega) = Y(\omega) \text{ for every } \omega \in \Omega \setminus Z$$

**Theorem 1.3.2** (Lindeberg-Lévy CLT). Let  $X_1, X_2, \dots$  be a sequence of i.i.d. random variables with  $\mathbb{E}X_1 = \mu$ ,  $\text{Var}(X_1) = \sigma^2 < \infty$ . Then the normalized partial sums

$$Z_N = \frac{S_N - \mathbb{E}S_N}{\sqrt{\text{Var}(S_N)}} = \frac{\sum_{i=1}^N X_i - N\mu}{\sigma\sqrt{N}} \rightarrow N(0, 1) \text{ in distribution}$$

where real random variables  $(Y_N)_{N=1}^\infty$  are said to **converge in distribution** to a random variable  $Y$  if their CDF's  $F_{Y_N}(t) := \mathbb{P}\{Y_N \leq t\}$ ,  $F_Y(t) := \mathbb{P}\{Y \leq t\}$  have

$$\lim_{N \rightarrow \infty} F_{Y_N}(t) = F_Y(t) \text{ for all } t \in \mathbb{R}$$

**Exercise 1.3.3.** Let  $X_1, X_2, \dots$  be a sequence of i.i.d. random variables with  $\mu, \sigma^2 < \infty$ . Show that

$$\mathbb{E} \left| \frac{1}{N} \sum_{i=1}^N X_i - \mu \right| = O\left(\frac{1}{\sqrt{N}}\right)$$

*Answer.* Considering the convex function  $\phi(x) = x^2$ , we can apply Jensen's to get

$$\left( \mathbb{E} \left| \frac{1}{N} \sum_{i=1}^N X_i - \mu \right| \right)^2 \leq \text{Var} \left( \frac{1}{N} \sum_{i=1}^N X_i \right) = \frac{\sigma^2}{N}$$

taking the square root of both sides yields the result. □

**Theorem 1.3.4** (Poisson limit theorem). Consider a sequence of  $N$ -tuples of independent random variables with entries  $X_{Ni}$  for  $1 \leq i \leq N$  with  $X_{Ni} \sim \text{Bernoulli}(p_{Ni})$ . Let  $S_N = \sum_{i=1}^N X_{Ni}$ , and suppose that as  $N \rightarrow \infty$

$$\max_{1 \leq i \leq N} p_{Ni} \rightarrow 0 \text{ and } \mathbb{E}S_N = \sum_{i=1}^N p_{Ni} \rightarrow \lambda$$

Then  $S_N \rightarrow \text{Poisson}(\lambda)$  in distribution, i.e. the CDF  $F_{S_N}(t) = \mathbb{P}\{S_N \leq t\}$  has

$$\lim_{N \rightarrow \infty} F_{S_N}(t) = \sum_{k=1}^{\lfloor t \rfloor} e^{-\lambda} \frac{\lambda^k}{k!}$$

## 2 Concentrations of sums of independent random variables

### 2.1 Why concentration inequalities?

Concentration inequalities quantify the variation of a random variable around its mean, and take the form

$$\mathbb{P}\{|X - \mu| \geq t\} \leq \text{something small}$$

**Proposition 2.1.2** (Tails of the normal distribution). Let  $Z \sim N(0, 1)$ . For  $t > 0$

$$\left(\frac{1}{t} - \frac{1}{t^3}\right) \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \leq \mathbb{P}\{Z \geq t\} \leq \frac{1}{t} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$$

In particular, for  $t \geq 1$ , the tail of  $Z$  has

$$\mathbb{P}\{Z \geq t\} \leq \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$$

More loosely, we can say

$$\mathbb{P}\{Z \geq T\} = \Theta\left(\frac{1}{te^{t^2/2}}\right) = \tilde{\Theta}\left(\frac{1}{e^{t^2/2}}\right)$$

*Proof.* For the upper bound, we substitute  $x = y + t$  to get

$$\int_t^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-y^2/2} e^{-ty} e^{-t^2/2} dy \leq \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \int_0^\infty e^{-ty} dy = \frac{1}{t} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$$

For the lower bound, we make use of the identity

$$\int_t^\infty e^{-x^2/2} dx \geq \int_t^\infty (1 - 3x^{-4}) e^{-x^2/2} dx = \left(\frac{1}{t} - \frac{1}{t^3}\right) e^{-t^2/2}$$

□

**Example 2.1.1.** Consider  $S_N = X_1 + \dots + X_N$ , each  $X_i \sim \text{Bernoulli}(1/2)$ . We have  $\mathbb{E}S_N = N/2$ ,  $\text{Var}(S_N) = N/4$ . From Chebyshev's inequality (Corollary 1.2.5), we get

$$\mathbb{P}\left\{\left|S_N - \frac{N}{2}\right| \geq \frac{N}{4}\right\} \leq \frac{4}{N} = O\left(\frac{1}{N}\right)$$

i.e. the probability of satisfying our concentration requirements goes to 0 linearly. Is this upper bound tight?

We know by the CLT (Theorem 1.3.2), that our normalized  $S_N$  converges in distribution to  $N(0, 1)$ . Then for large  $N$ , we should see that

$$\mathbb{P}\left\{\left|S_N - \frac{N}{2}\right| \geq \frac{N}{4}\right\} = \mathbb{P}\left\{\left|\frac{S_N - \frac{N}{2}}{\sqrt{\frac{N}{4}}}\right| \geq \sqrt{\frac{N}{4}}\right\} \approx \mathbb{P}\left\{|Z| \geq \sqrt{\frac{N}{4}}\right\} \leq \frac{1}{\sqrt{2\pi N}} e^{-N/8} = \tilde{O}\left(\frac{1}{e^{N/8}}\right)$$

which is exponentially fast (by Proposition 2.1.2). However this central limit theorem argument can't be made rigorous, since the error in approximating normalized  $S_N$  with  $Z$  decays too slowly (in fact slower than linearly via Theorem 2.1.3). It turns out that for these sums, we get light tails much faster than we approximate  $N(0, 1)$ .

**Theorem 2.1.3** (Berry-Esseen CLT). In the setting of Theorem 1.3.2, for all  $N$

$$|F_{Z_N}(t) - F_Z(t)| \leq \frac{\rho}{\sqrt{N}} \quad \text{for all } t \in \mathbb{R}$$

where  $\rho = \mathbb{E}|X_1 - \mu|^3 / \sigma^3$ .

Note that in comparison to Theorem 1.3.2 it additionally requires the third moment  $\mathbb{E}X_1^3 < \infty$ , and in turn provides a *quantitative* rate for *uniform* convergence in distribution to  $N(0, 1)$ .

**Exercise 2.1.4** (Truncated normal distribution). Let  $Z \sim N(0, 1)$ . Show that for all  $t > 0$

$$\mathbb{E}Z^2 \mathbb{1}_{\{Z \geq t\}} = t \frac{1}{\sqrt{2\pi}} e^{-t^2/2} + \mathbb{P}(Z \geq t) \leq \left(t + \frac{1}{t}\right) \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$$

*Answer.* To prove the equality

$$\mathbb{E}Z^2 \mathbb{1}_{\{Z \geq t\}} := \int_t^\infty z^2 \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \frac{1}{\sqrt{2\pi}} \left( \left[-ze^{-z^2/2}\right]_t^\infty + \int_t^\infty e^{-z^2/2} dz \right) = t \frac{1}{\sqrt{2\pi}} e^{-t^2/2} + \mathbb{P}\{Z \geq t\}$$

The inequality follows from the tail upper bound from Proposition 2.1.2.

□



## 2.2 Hoeffding's inequality

**Definition 2.2.1** (Rademacher distribution). We say a random variable  $X$  has **Rademacher** distribution if

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

**Theorem 2.2.2** (Hoeffding's inequality). Let  $X_1, \dots, X_N$  be independent Rademacher random variables. Let  $a = (a_1, \dots, a_n) \in \mathbb{R}^N$ . For  $t \geq 0$

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

*Proof.* WLOG assume  $\|a\|_2^2 = 1$ . If we prove this version of the theorem, then for any  $b = ca \in \mathbb{R}^N$ ,

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

We apply Markov's inequality to the MGF of  $\sum_{i=1}^N a_i X_i$

$$\mathbb{P}\left\{\sum_{i=1}^N a_i X_i \geq t\right\} = \mathbb{P}\left\{\exp\left(\lambda \sum_{i=1}^N a_i X_i\right) \geq \exp(\lambda t)\right\} \leq \frac{\mathbb{E} \exp\left(\lambda \sum_{i=1}^N a_i X_i\right)}{\exp(\lambda t)}$$

Examining the numerator of the right side of the inequality

$$\begin{aligned} \mathbb{E} \exp\left(\lambda \sum_{i=1}^N a_i X_i\right) &= \prod_{i=1}^N \mathbb{E} \exp(\lambda a_i X_i) && \text{(by independence of } X_i) \\ &= \prod_{i=1}^N \cosh(\lambda a_i) && \text{(by definition of } \mathbb{E} \text{ of Rademacher RVs)} \\ &\leq \prod_{i=1}^N \exp(\lambda^2 a_i^2 / 2) && \text{(by Exercise 2.2.3)} \\ &= \exp(\lambda^2 / 2) && \text{(since } \|a\|_2^2 = 1) \end{aligned}$$

To complete the proof we optimize  $\lambda$  to minimize the right hand side of the obtained tail bound inequality,  $\exp(\lambda^2/2 - \lambda t)$ . Setting  $d(\lambda^2/2 - \lambda t)/d\lambda = \lambda - t = 0$  yields the minimum  $\lambda = t$ . This yields the desired inequality

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

□

**Exercise 2.2.3** (Bounding the hyperbolic cosine). Show that

$$\cosh(x) \leq \exp(x^2/2) \quad \text{for all } x \in \mathbb{R}$$

*Answer.* Recalling that  $e^x = \sum_{k=0}^{\infty} x^k/k!$  for all  $x \in \mathbb{R}$ , we can compute Taylor expansions that converge on  $\mathbb{R}$

$$\begin{aligned} \cosh(x) &= \frac{1}{2} \left( \sum_{k=0}^{\infty} \frac{x^k}{k!} + \sum_{k=0}^{\infty} \frac{(-x)^k}{k!} \right) = \sum_{k=0}^{\infty} \frac{x^{2k}}{(2k)!} \\ e^{x^2/2} &= \sum_{k=0}^{\infty} \frac{(x^2/2)^k}{k!} = \sum_{k=0}^{\infty} \frac{x^{2k}}{k!2^k} \end{aligned}$$

Note that for  $k = 0$ , the terms match. For  $k \geq 1$ ,

$$\frac{x^{2k}}{(2k)!} \leq \frac{x^{2k}}{k!2^k} \iff (2k)! \geq k!2^k \iff 2k \cdot \dots \cdot k+1 \geq \underbrace{2 \cdot \dots \cdot 2}_{k \text{ times}}$$

where the last statement holds if  $k+1 \geq 2 \iff k \geq 1$ . Hence for all  $x \in \mathbb{R}$ , the partial sums of the expansion of  $\cosh(x)$  are upper bounded by the partial sums of the expansion of  $e^{x^2/2}$ , which implies the same for their limits. □

**Remark 2.2.4.** We can use Hoeffding's to analyze the  $N$  coin flips from Example 2.1.1, achieving the desired (non-asymptotic) exponentially decaying tail probabilities.

$$\mathbb{P} \left\{ \sum_{i=1}^N X_i \geq 3N/4 \right\} = \mathbb{P} \left\{ \sum_{i=1}^N 2X_i - 1 \geq N/2 \right\} \leq \exp(-(N/2)^2/2N) = \exp(-N/8)$$

**Theorem 2.2.6** (Hoeffding's inequality for general bounded RVs). Let  $X_1, \dots, X_N$  be independent random variables, with each  $X_i$ 's support  $[m_i, M_i]$ . For  $t > 0$

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

**Exercise 2.2.7.** Prove Theorem 2.2.6, possibly with some absolute constant instead of 2 in the tail.

*Answer.* We consider  $X_i$  with mean 0. For  $X_i$  without mean 0, we set  $Y_i = X_i - \mathbb{E}X_i$  and proceed in the proof with  $Y_i$  which have the same support length as the  $X_i$ . The argument follows as in the proof of Theorem 2.2.2 differing only at the part where we obtain a bound for the MGF of the individual  $X_i$ .

**Claim** (Hoeffding's lemma). For a bounded random variable  $X \in [m, M]$  with mean 0, we have

$$\mathbb{E} \exp(\lambda X) \leq \exp(\lambda^2(M - m)^2/2)$$

With the claim we arrive at the inequality

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

Optimizing  $\lambda$  yields  $\lambda = t / \sum_{i=1}^N (M_i - m_i)^2$  and we get

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

Note that the  $1/2$  in the tail is looser than the 2 in the theorem statement, which we can get if we prove a tighter version of Hoeffding's lemma (with  $/8$  instead of  $/2$ ) using Taylor approximations. For the proof of the claim, we follow an argument by symmetrization presented in the proof of Lemma 5 from these CS229 lecture notes<sup>¶</sup> by John Duchi. Consider  $X'$  an independent copy of  $X$ . We have

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

Where the second inequality is by Jensen's. Since  $X - X'$  has a symmetric distribution, it has the same distribution as  $S(X - X')$ , where  $S$  is a Rademacher random variable, giving us the last equality.

$$\begin{aligned} \mathbb{E}_X \mathbb{E}_{X', S} \exp(\lambda S(X - X')) &= \mathbb{E}_{X, X'} \cosh(\lambda(X - X')) && \text{(by definition of } \mathbb{E}_S \text{ of Rademacher RVs)} \\ &\leq \mathbb{E}_{X, X'} \exp(\lambda^2(X - X')^2/2) && \text{(by Exercise 2.2.3)} \\ &\leq \exp(\lambda^2(M - m)^2/2) && \text{(since } |X - X'| \leq |M - m|) \end{aligned}$$

□

**Exercise 2.2.8** (Boosting randomized algorithms). Suppose we have a randomized algorithm for a decision problem that is correct with probability  $1/2 + \varepsilon$  for some  $\varepsilon > 0$ . Show that running the algorithm  $N$  times independently and taking the majority yields the correct answer with probability  $\geq 1 - \delta$  for  $N \geq (1/2\varepsilon^2) \log \frac{1}{\delta}$ .

*Answer.*

□

**Exercise 2.2.9** (Robust mean estimation). Suppose we want to estimate the mean  $\mu$  of a random variable  $X$  from  $X_1, \dots, X_N$  copies drawn independently. We want an  $\varepsilon$ -accurate estimate (falls within  $(\varepsilon - \mu, \varepsilon + \mu)$ ).

(a) Show a sample size  $N = O(\sigma^2/\varepsilon^2)$  is sufficient for an  $\varepsilon$ -accurate estimate w.p.  $\geq 3/4$ , where  $\sigma^2 = \text{Var}(X)$ .

<sup>¶</sup><http://cs229.stanford.edu/extra-notes/hoeffding.pdf>

*Answer.*

□

- (b) Show a sample size  $N = O(\log(1/\delta)\sigma^2/\varepsilon^2)$  is sufficient for an  $\varepsilon$ -accurate estimate w.p.  $\geq 1 - \delta$ .

*Answer.*

□

**Exercise 2.2.10** (Small ball probabilities). Let  $X_1, \dots, X_N$  be non-negative independent random variables with continuous distributions. Assume their densities are bounded by 1.

- (a) Show that the MGF of  $X_i$  satisfies

$$\mathbb{E} \exp(-tX_i) \leq \frac{1}{t} \quad \text{for all } t \geq 0$$

*Answer.*

□

- (b) Deduce that for any  $\varepsilon > 0$  we have

$$\mathbb{P} \left\{ \sum_{i=1}^N X_i \leq \varepsilon N \right\} \leq (e\varepsilon)^N$$

*Answer.*

□

## 2.3 Chernoffs's inequality