Stat 6390.001: High Dimensional Probability and Statistics, Fall 2026 **Topic 1**

High Dimensional Probability and Statistics

Lecturer: Yun Wei *Scribe: Students in Fall 2026*

Contents

Chapter 1

Basic tail and concentration bounds

1.1 Introduction

A central challenge in high-dimensional statistics is the *curse of dimensionality*. As the dimension d grows, the sample space expands rapidly, so a fixed number of observations becomes increasingly sparse relative to the size of the space. As a result, methods and intuition that work well in low dimensions may become much less informative in high-dimensional settings.

High-dimensional problems also exhibit geometric and probabilistic behavior that can differ substantially from low-dimensional intuition. For example, if

$$X \sim N(0, I_d),$$

then the density is maximized at the origin, but a typical sample is not concentrated near the origin. Instead,

$$\|X\|_2 \approx \sqrt{d},$$

so most samples lie near a thin shell of radius approximately $\sqrt{d}$.

These phenomena motivate the study of probabilistic tools that remain useful when the dimension is large. In particular, this lecture develops tail and concentration bounds that quantify how likely a random variable is to deviate from its mean, with an emphasis on non-asymptotic guarantees.

1.1.1 One-dimensional and multivariate CLT

Let $X_1, \ldots, X_n$ be i.i.d. random variables with mean μ and finite variance σ^2. The one-dimensional CLT states that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right) \xrightarrow{d} N(0, \sigma^2) \qquad \text{as } n \to \infty. \tag{1.1}$$

The multivariate version has the same form. Suppose now that $X_1, \ldots, X_n$ are i.i.d. random vectors in $\mathbb{R}^d$ with

$$\mathbb{E}[X_i] = \mu \in \mathbb{R}^d, \qquad \mathrm{Cov}(X_i) = \Sigma.$$

Then

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right) \xrightarrow{d} N(0, \Sigma) \qquad \text{as } n \to \infty. \tag{1.2}$$

For a d-dimensional Gaussian vector $X \sim N(\mu, \Sigma)$ with positive definite covariance matrix Σ, the density is

$$f(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}. \tag{1.3}$$

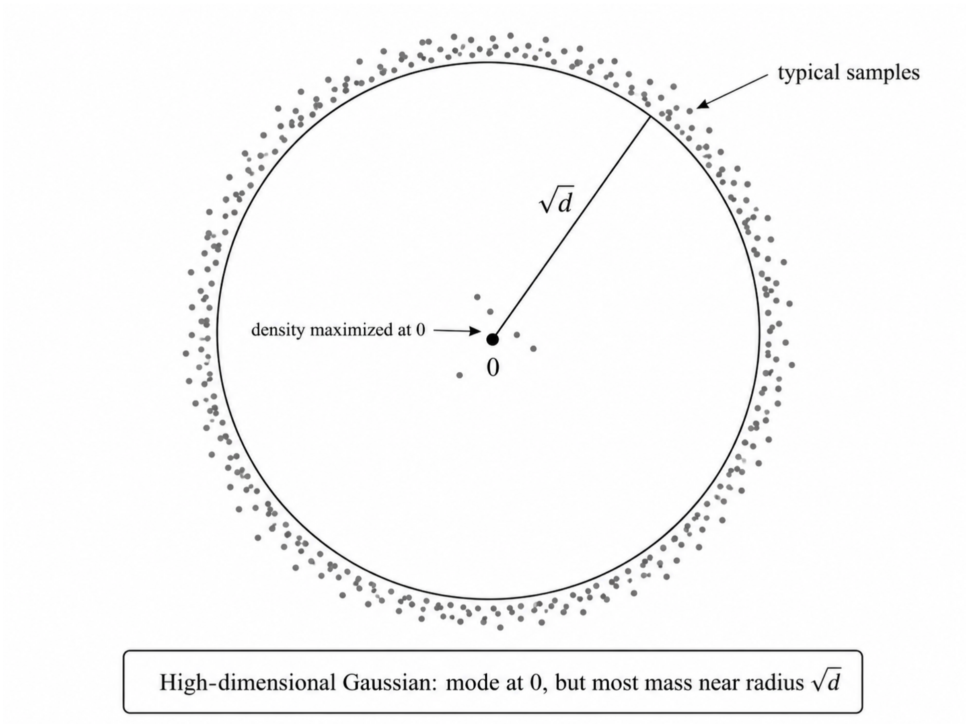

Figure 1.1: A schematic illustration of a high-dimensional standard Gaussian distribution.

Example 1.1.1 (A First High-Dimensional Phenomenon)**.** Consider the standard Gaussian case $X \sim N(0, I_d)$. The density is maximized at the origin, so the mode is $x = 0$. However, when d is large, a typical draw is not close to the origin. Indeed,

$$\mathbb{E}\|X\|_2^2 = \mathbb{E}\left[\sum_{j=1}^{d} X_j^2\right] = \sum_{j=1}^{d} \mathbb{E}[X_j^2] = d. \tag{1.4}$$

Thus the typical Euclidean radius is of order $\sqrt{d}$; more precisely, $\|X\|_2^2$ is a chi-square random variable with d degrees of freedom and is concentrated around d. Hence most of the Gaussian probability mass lies in a thin shell with radius approximately $\sqrt{d}$, even though the density itself is largest at the origin, shown in Figure 1. This illustrates a basic feature of high-dimensional geometry: the location of the mode need not describe where a typical sample lies.

1.2 Classical Tail Bounds

The purpose of a tail bound is to control probabilities such as $\mathbb{P}(X \geq t)$ or $\mathbb{P}(|X - \mathbb{E}X| \geq t)$. The amount of information available about the distribution determines the quality of the bound. Markov's inequality uses only a first moment; Chebyshev's inequality uses a second moment; and the Chernoff method exploits the moment generating function (MGF), thereby incorporating information from all moments when the MGF exists [1].

1.2.1 Markov and Chebyshev inequalities

Proposition 1.2.1 (Markov's inequality)**.** *If random variable X is nonnegative and $\mathbb{E}[X] < \infty$, then for* *every $t > 0$,*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}. \tag{1.5}$$

Proposition 1.2.2 (Chebyshev's inequality)**.** *If random variable X has finite variance and $\mu = \mathbb{E}[X]$, then for every $t > 0$,*

$$\mathbb{P}\big(|X - \mu| \geq t\big) \leq \frac{\operatorname{Var}(X)}{t^2}. \tag{1.6}$$

Chebyshev's inequality is an immediate consequence of Markov's inequality applied to the nonnegative random variable $(X - \mu)^2$. More generally, if the kth centered absolute moment exists, then

$$\mathbb{P}\big(|X - \mu| \geq t\big) = \mathbb{P}\big(|X - \mu|^k \geq t^k\big) \leq \frac{\mathbb{E}|X - \mu|^k}{t^k}. \tag{1.7}$$

1.2.2 From Markov to the Chernoff bound

The exponential function is particularly effective because it converts sums into products and leads naturally to the moment generating function.

Lemma 1.2.3 (Chernoff bound)**.** *Let $\mu = \mathbb{E}[X]$, and suppose*

$$\phi(\lambda) = \mathbb{E}\big[e^{\lambda(X-\mu)}\big]$$

exists for $\lambda \in [0, b]$. Then for every $t > 0$,

$$\log \mathbb{P}(X - \mu \geq t) \leq \inf_{\lambda \in [0,b]} \left\{\log \phi(\lambda) - \lambda t\right\}. \tag{1.8}$$

Proof. Fox $\lambda > 0$. Since the exponential function is increasing,

$$\mathbb{P}(X - \mu \geq t) = \mathbb{P}\big(\lambda(X - \mu) \geq \lambda t\big) \tag{1.9}$$

$$= \mathbb{P}\big(e^{\lambda(X-\mu)} \geq e^{\lambda t}\big). \tag{1.10}$$

Applying Markov's inequality to the nonnegative random variable $e^{\lambda(X-\mu)}$ gives

$$\mathbb{P}(X - \mu \geq t) \leq \frac{\mathbb{E}[e^{\lambda(X-\mu)}]}{e^{\lambda t}} = \phi(\lambda)e^{-\lambda t}. \tag{1.11}$$

Taking logarithms yields

$$\log \mathbb{P}(X - \mu \geq t) \leq \log \phi(\lambda) - \lambda t.$$

Because the inequality holds for every admissible λ, minimizing the right-hand side over $\lambda \in [0, b]$ gives the result. □

Remark 1.2.4. The optimization over λ is essential. Different values of λ produce different valid bounds, and the infimum selects the sharpest one obtained from this argument. In applications, the main work is therefore to control the MGF and then solve a one-dimensional optimization problem.

1.3 Sub-Gaussian Random Variables

1.3.1 Gaussian Tail Bound

As a first example, consider a Gaussian random variable

$$X \sim N(\mu, \sigma^2).$$

Its moment generating function is

$$\mathbb{E}[e^{\lambda X}] = \exp\left(\mu\lambda + \frac{\sigma^2\lambda^2}{2}\right), \qquad \lambda \in \mathbb{R}. \tag{1.12}$$

Applying the preceding lemma, the upper-tail bound is determined by

$$\inf_{\lambda>0}\left\{\log \mathbb{E}[e^{\lambda(X-\mu)}] - \lambda t\right\} = \inf_{\lambda>0}\left\{\frac{\sigma^2\lambda^2}{2} - \lambda t\right\}. \tag{1.13}$$

Here we used the fact that

$$X - \mu \sim N(0, \sigma^2).$$

The quadratic expression is minimized at

$$\lambda^* = \frac{t}{\sigma^2},$$

which gives

$$\inf_{\lambda>0}\left\{\frac{\sigma^2\lambda^2}{2} - \lambda t\right\} = -\frac{t^2}{2\sigma^2}. \tag{1.14}$$

Therefore, for every $t > 0$,

$$\mathbb{P}(X - \mu \geq t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right). \tag{1.15}$$

(Normal distribution is light tail)

1.3.2 Definition and basic tail bounds

Definition 1.3.1 (Sub-Gaussian random variable)**.** A random variable X with mean $\mu = \mathbb{E}[X]$ is *sub-Gaussian with parameter* $\sigma > 0$ if

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq \exp\left(\frac{\sigma^2\lambda^2}{2}\right) \qquad \text{for all } \lambda \in \mathbb{R}. \tag{1.16}$$

Remark 1.3.2. If X is σ-sub-Gaussian for some $\sigma > 0$, then X is also σ'-sub-Gaussian for some $\sigma' \geq \sigma$.

The definition says that the centered MGF of X grows no faster than the centered MGF of a Gaussian $N(0, \sigma^2)$. Therefore, the same Chernoff calculation gives the Gaussian-type upper-tail bound

$$\mathbb{P}(X - \mu \geq t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right), \qquad t > 0. \tag{1.17}$$

Lemma 1.3.3 (Symmetry of sub-Gaussianity)**.** *A random variable X is sub-Gaussian with parameter σ if and only if $-X$ is sub-Gaussian with the same parameter.*

Proof. Let $\mu = \mathbb{E}[X]$. Recall that X is sub-Gaussian with parameter σ if

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq \exp\left(\frac{\sigma^2\lambda^2}{2}\right), \qquad \lambda \in \mathbb{R}. \tag{1.18}$$

Suppose first that X is sub-Gaussian. Since $\mathbb{E}[-X] = -\mu$,

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda((-X)-\mathbb{E}[-X])}\right] &= \mathbb{E}\left[e^{-\lambda(X-\mu)}\right] && (1.19)\\
&\leq \exp\left(\frac{\sigma^2(-\lambda)^2}{2}\right) && (1.20)\\
&= \exp\left(\frac{\sigma^2\lambda^2}{2}\right). && (1.21)
\end{aligned}$$

Hence $-X$ is sub-Gaussian with parameter σ.

Conversely, suppose that $-X$ is sub-Gaussian. Then

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] = \mathbb{E}\left[e^{-\lambda((-X)-\mathbb{E}[-X])}\right] \tag{1.22}$$

$$\leq \exp\left(\frac{\sigma^2(-\lambda)^2}{2}\right) \tag{1.23}$$

$$= \exp\left(\frac{\sigma^2\lambda^2}{2}\right). \tag{1.24}$$

Therefore X is sub-Gaussian with parameter σ. □

Lemma 1.3.4 (Lower Tail)**.** *If X is sub-Gaussian with parameter σ, then for every $t > 0$,*

$$\mathbb{P}(X - \mathbb{E}[X] \leq -t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right). \tag{1.25}$$

Proof. By the symmetry lemma, $-X$ is also sub-Gaussian with parameter σ. Moreover,

$$\mathbb{P}(X - \mathbb{E}[X] \leq -t) = \mathbb{P}(-X - \mathbb{E}[-X] \geq t) \tag{1.26}$$

$$\leq \exp\left(-\frac{t^2}{2\sigma^2}\right), \tag{1.27}$$

where the last step follows from the upper-tail bound applied to $-X$. □

Now if we combine upper and lower tails with the union bound yields the two-sided concentration inequality we will get the following

Lemma 1.3.5. *If X is sub-Gaussian with parameter $\sigma > 0$,*

$$\mathbb{P}\big(|X - \mathbb{E}X| \geq t\big) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right), \qquad t > 0. \tag{1.28}$$

Note: This much faster decay is one reason sub-Gaussian assumptions are central in non-asymptotic high-dimensional statistics [1].

Example 1.3.6 (Rademacher random variable)**.** A Rademacher random variable ε takes values in $\{-1, +1\}$ equiprobably:

In particular, $\mathbb{E}[\varepsilon] = 0$. Its MGF is

$$\mathbb{E}[e^{\lambda\varepsilon}] = \frac{1}{2}e^{\lambda} + \frac{1}{2}e^{-\lambda} \tag{1.29}$$

$$= \frac{1}{2}\left(\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} + \sum_{k=0}^{\infty}\frac{(-\lambda)^k}{k!}\right) \tag{1.30}$$

$$= \sum_{m=0}^{\infty}\frac{\lambda^{2m}}{(2m)!}, \tag{1.31}$$

where the odd-order terms cancel.

Since $(2m)! \geq 2^m m!$,

$$\mathbb{E}[e^{\lambda\varepsilon}] \leq \sum_{m=0}^{\infty}\frac{\lambda^{2m}}{2^m m!} \tag{1.32}$$

$$= e^{\lambda^2/2}. \tag{1.33}$$

Since $\mathbb{E}[\varepsilon]=0$, it follows that

$$\mathbb{E}[e^{\lambda(\varepsilon-\mathbb{E}[\varepsilon])}]\le e^{\lambda^2/2}.$$

Therefore, ε is sub-Gaussian with parameter $\sigma=1$.

This example shows that sub-Gaussianity is not the same as Gaussianity. A Rademacher variable is discrete and bounded, yet its MGF is dominated by that of a standard Gaussian. The sub-Gaussian class is therefore broad enough to include many non-Gaussian random variables while retaining Gaussian-like tail behavior.

Note:"equiprobably" meaning

$$\mathbb{P}(\varepsilon=1)=\mathbb{P}(\varepsilon=-1)=\frac{1}{2}. \tag{1.34}$$

Therefore,

$$\mathbb{E}[\varepsilon]=\frac{1}{2}(1)+\frac{1}{2}(-1)=0. \tag{1.35}$$

1.3.3 Example of Sub-Gaussian Variables

Example 1.3.7 (Bounded random variables)**.** Let X be zero-mean, and supported on some interval $[a,b]$. We will show that X is sub-Gaussian with parameter at most $\sigma=b-a$.

Letting X' be an independent copy of X, so that $X\overset{d}{=}X'$ and $X\perp\!\!\!\perp X'$. For any $\lambda\in\mathbb{R}$, we have

$$\mathbb{E}_X\left[e^{\lambda X}\right]=\mathbb{E}_X\left[e^{\lambda(X-\mathbb{E}_{X'}[X'])}\right]\overset{\text{(i)}}{\le}\mathbb{E}_X\left[\mathbb{E}_{X'}e^{\lambda(X-X')}\right]=\mathbb{E}_{X,X'}\left[e^{\lambda(X-X')}\right]$$

The first equality is because $\mathbb{E}[X']=0$, and (i) follows from the the convexity of the exponential, and Jensen's inequality. Letting ε be an independent Rademacher variable, note that since (X,X') are i.i.d, $X-X'\overset{d}{=}X'-X$, so $(X-X')$ is symmetric about 0. Thus, the distribution of $(X-X')$ is the same as that of $\varepsilon(X-X')$, so that we have

$$\mathbb{E}_{X,X'}\left[e^{\lambda(X-X')}\right]=\mathbb{E}_{X,X'}\left[\mathbb{E}_\varepsilon\left[e^{\lambda\varepsilon(X-X')}\right]\right]\overset{\text{(ii)}}{\le}\mathbb{E}_{X,X'}\left[e^{\frac{\lambda^2}{2}(X-X')^2}\right]\le e^{\frac{\lambda^2}{2}(b-a)^2}$$

The first equality follows from Tonelli's theorem and (ii) follows from the sub-Gaussianity of Radamacher variable and last inequality is because $|X-X'|\le b-a$. Putting everything together, we have shown that X is sub-Gaussian with parameter at most $\sigma=b-a$.

Remark 1.3.8. This result is useful, but it can be sharpened. It can be shown that X is sub-Gaussian with parameter at most $\sigma=\frac{b-a}{2}$ (Exercise 2.4). The technique used in this example is a simple example of a *symmetrization argument*, in which we first introduce an independent copy X', and then symmetrize the problem with a Rademacher variable. Such symmetrization arguments are useful in a variety of contexts.

1.3.4 Hoeffding's Inequality

Another property of sub-Gaussianity is preserved by linear operations.

Lemma 1.3.9. *Let X_1 and X_2 be independent sub-Gaussian variables with parameters σ_1 and σ_2, respectively, then X_1+X_2 is sub-Gaussian with parameter $\sqrt{\sigma_1^2+\sigma_2^2}$.*

Proof. By independence we have

$$\mathbb{E}\left[e^{\lambda(X_1+X_2-\mathbb{E}X_1-\mathbb{E}X_2)}\right]=\mathbb{E}\left[e^{\lambda(X_1-\mathbb{E}X_1)}\right]\mathbb{E}\left[e^{\lambda(X_2-\mathbb{E}X_2)}\right]\le e^{\frac{\lambda^2}{2}\sigma_1^2}e^{\frac{\lambda^2}{2}\sigma_2^2}=e^{\frac{\lambda^2}{2}(\sigma_1^2+\sigma_2^2)}$$

$\square$

As a consequence of this fact, we obtain an important result known as the *Hoeffding's inequality*:

Proposition 1.3.10 (Hoeffding's inequality)**.** *Suppose that the variables X_i, $i = 1, \dots, n$, are independent, and X_i has mean μ_i and sub-Gaussian parameter σ_i. Then for all $t \geq 0$, we have*

$$\mathbb{P}\left[\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right] \leq \exp\left\{-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2}\right\}$$

and

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

Proof. By previous lemma, $\sum_{i=1}^{n} X_i$ is $\sqrt{\sigma_1^2 + \cdots + \sigma_n^2}$ sub-Gaussian. So, by basic sub-Gaussian tail bound we have the results. $\square$

Corollary 1.3.11. *If X_i, $i = 1, \dots, n$, are i.i.d, with mean μ and sub-Gaussian parameter σ. Then for all $t \geq 0$, we have*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq t\right] \leq 2\exp\left\{-\frac{nt^2}{2\sigma^2}\right\}$$

and

$$\mathbb{P}\left[\sqrt{n}\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq t\right] \leq 2\exp\left\{-\frac{t^2}{2\sigma^2}\right\}$$

Remark 1.3.12. The above can be viewed as a finite sample Central Limit Theorem. CLT states that when $n \to \infty$,

$$\sqrt{n}\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \to \mathcal{N}(0, \sigma^2).$$

The second statement of the above corollary is a finite sample version of CLT, by stating that the tail of $\sqrt{n}\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right|$ is upper bounded by the tail of $\mathcal{N}(0, \sigma^2)$.

Proof. By Hoeffding,

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

Second statement follows analogously. $\square$

One application of Hoeffding's bound is to derive finite sample guarantees for the error of the sample mean.

Example 1.3.13 (Finite-sample error guarantee)**.** Let $\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ and fix a confidence level $\delta \in (0, 1)$. By the preceding corollary,

$$\mathbb{P}\left[\left|\overline{X}_n - \mu\right| \geq t\right] \leq 2\exp\left\{-\frac{nt^2}{2\sigma^2}\right\}$$

Choose $t = \sqrt{\frac{2\sigma^2}{n}\log\left(\frac{2}{\delta}\right)}$, so that $\delta = 2\exp\left\{-\frac{nt^2}{2\sigma^2}\right\}$. In other words, with probability at least $1 - \delta$,

$$\left|\overline{X}_n - \mu\right| \leq \sqrt{\frac{2\sigma^2}{n}\log\left(\frac{2}{\delta}\right)}$$

Example 1.3.14 (Expected absolute error)**.** We can also control the expected absolute error $\mathbb{E}\left|\overline{X}_n - \mu\right|$ directly from the tail bound. Recall that for any nonnegative random variable Y, Tonelli's theorem gives

$$\mathbb{E}[Y] = \int_0^\infty \mathbb{P}(Y \geq t)\, dt$$

Taking $Y = \left|\overline{X}_n - \mu\right|$, the preceding corollary and the trivial bound $\mathbb{P}(Y \geq t) \leq 1$ imply

$$\mathbb{P}\left[\left|\overline{X}_n - \mu\right| \geq t\right] \leq \min\left\{1,\ 2\exp\left\{-\frac{nt^2}{2\sigma^2}\right\}\right\}$$

Let t_0 be the point at which the two upper bounds agree. Thus $2e^{-\frac{nt_0^2}{2\sigma^2}} = 1$, which gives $t_0 = \sigma\sqrt{\frac{2\log 2}{n}}$. Hence

$$\begin{aligned}
\mathbb{E}\left|\overline{X}_n - \mu\right| &= \int_0^\infty \mathbb{P}\left[\left|\overline{X}_n - \mu\right| \geq t\right]\, dt \\
&\leq \int_0^{t_0} 1\, dt + \underbrace{\int_{t_0}^\infty 2\exp\left\{-\frac{nt^2}{2\sigma^2}\right\} dt}_{z = \frac{\sqrt{n}}{\sigma}t, \quad dt = \frac{\sigma}{\sqrt{n}}\, dz} \\
&= t_0 + \frac{2\sigma}{\sqrt{n}} \int_{\frac{\sqrt{n}}{\sigma}t_0}^\infty e^{-z^2/2}\, dz \\
&= \frac{\sigma}{\sqrt{n}}\left(\sqrt{2\log 2} + 2\int_{\sqrt{2\log 2}}^\infty e^{-z^2/2}\, dz\right) \\
&\leq 2\frac{\sigma}{\sqrt{n}}.
\end{aligned}$$

We conclude our discussion of sub-Gaussianity with a theorem that provides different characterizations of sub-Gaussian variables.

Theorem 1.3.15 (Equivalent characterizations of sub-Gaussian variables)**.** *Given any zero-mean random variable X, the following properties are equivalent:*

1. *There is a constant $\sigma \geq 0$ such that*

$$\mathbb{E}\left[e^{\lambda X}\right] \leq e^{\frac{\lambda^2\sigma^2}{2}} \qquad \text{for all } \lambda \in \mathbb{R}.$$

2. *There is a constant $c \geq 0$ and Gaussian random variable $Z \sim \mathcal{N}(0, \tau^2)$ such that*

$$\mathbb{P}\left[|X| \geq s\right] \leq c\,\mathbb{P}\left[|Z| \geq s\right] \qquad \text{for all } s \geq 0.$$

3. *There is a constant $\theta \geq 0$ such that*

$$\mathbb{E}\left[X^{2k}\right] \leq \frac{(2k)!}{2^k k!}\theta^{2k} \qquad \text{for all } k = 1, 2, \ldots$$

4. *There is a constant $\sigma \geq 0$ such that*

$$\mathbb{E}\left[e^{\frac{\lambda X^2}{2\sigma^2}}\right] \leq \frac{1}{\sqrt{1-\lambda}} \qquad \text{for all } \lambda \in [0, 1).$$

Proof. See Appendix A (Section 2.4) in [1]. □

1.4 Symmetrization

Lemma 1.4.1. *Let $Y \perp\!\!\!\perp Z$ with $\mathbb{E}[Z] = 0$. Then for any convex function F,*

$$\mathbb{E}\left[F(Y)\right] \leq \mathbb{E}\left[F(Y+Z)\right]. \tag{1.36}$$

Example 1.4.2. The function $F(x) = e^{\lambda x}$ is convex. Suppose $\mathbb{E}[X] = 0$. Applying Lemma 1.4.1 with $Y = X$ and $Z = -X'$, where $X' \perp\!\!\!\perp X$ and $X' \sim \mathrm{Law}(X)$, we obtain

$$\mathbb{E}_X\left[e^{\lambda X}\right] \leq \mathbb{E}_{X,X'}\left[e^{\lambda(X-X')}\right]. \tag{1.37}$$

Proof of Lemma 1.4.1. For every $y \in \mathbb{R}$, using $\mathbb{E}[Z] = 0$ and Jensen's inequality,

$$F(y) = F\left(y + \mathbb{E}_Z[Z]\right) = F\left(\mathbb{E}_Z\left[y+Z\right]\right) \leq \mathbb{E}_Z\left[F(y+Z)\right].$$

Substituting $y = Y$ gives $F(Y) \leq \mathbb{E}_Z\left[F(Y+Z)\right]$, and taking expectations over Y yields

$$\mathbb{E}_Y\left[F(Y)\right] \leq \mathbb{E}_Y\mathbb{E}_Z\left[F(Y+Z)\right] = \mathbb{E}_{Y,Z}\left[F(Y+Z)\right],$$

where the last equality uses $Y \perp\!\!\!\perp Z$. $\square$

Definition 1.4.3. A random variable X is called *symmetric* if it has the same distribution as $-X$.

Remark 1.4.4. If X has pdf $f(x)$, then X is symmetric if and only if $f(x) = f(-x)$.

Lemma 1.4.5 (Constructing symmetric distributions)**.** *Let X be a random variable and let ε be an independent Rademacher random variable.*

(a) εX and $\varepsilon\,|X|$ are identically distributed and symmetric.

(b) If X is symmetric, then both εX and $\varepsilon\,|X|$ have the same distribution as X.

(c) If X' is an independent copy of X, then $X - X'$ is symmetric.

Proof. We only prove that εX is symmetric. For every event $A \subseteq \mathbb{R}$, conditioning on the value of ε gives

$$\mathbb{P}\left[\varepsilon X \in A\right] = \tfrac{1}{2}\mathbb{P}\left[X \in A\right] + \tfrac{1}{2}\mathbb{P}\left[-X \in A\right] = \tfrac{1}{2}\mathbb{P}\left[-X \in A\right] + \tfrac{1}{2}\mathbb{P}\left[X \in A\right] = \mathbb{P}\left[-\varepsilon X \in A\right]. \quad \square$$

Lemma 1.4.6 (Symmetrization)**.** *Let $X_1, \ldots, X_N$ be independent, mean-zero random vectors in some normed space, and let $\varepsilon_1, \ldots, \varepsilon_N$ be independent Rademacher random variables. Then*

$$\frac{1}{2}\,\mathbb{E}\left\|\sum_{i=1}^N \varepsilon_i X_i\right\| \;\leq\; \mathbb{E}\left\|\sum_{i=1}^N X_i\right\| \;\leq\; 2\,\mathbb{E}\left\|\sum_{i=1}^N \varepsilon_i X_i\right\|. \tag{1.38}$$

Proof. Let $(X_1', \ldots, X_N')$ be independent copies of $(X_1, \ldots, X_N)$. By Lemma 1.4.5(c) the vectors $X_i - X_i'$ are independent and symmetric (this step uses the assumption that X_i are independent), so that

$$(X_1 - X_1', \ldots, X_N - X_N') \stackrel{d}{=} \left(\varepsilon_1(X_1 - X_1'), \ldots, \varepsilon_N(X_N - X_N')\right). \tag{1.39}$$

Upper bound. Set $p := \mathbb{E}\left\|\sum_{i=1}^N X_i\right\|$. Since $\mathbb{E}\left[\sum_{i=1}^N X_i'\right] = 0$, we may apply Lemma 1.4.1 with $F(x) = \|x\|$, $Y = \sum_{i=1}^N X_i$ and $Z = -\sum_{i=1}^N X_i'$ to get

$$p = \mathbb{E}\left\|\sum_{i=1}^N X_i - \mathbb{E}\left[\sum_{i=1}^N X_i'\right]\right\| \leq \mathbb{E}\left\|\sum_{i=1}^N (X_i - X_i')\right\|.$$

Then by (1.39) and the triangle inequality,

$$\begin{aligned}\mathbb{E}\left\|\sum_{i=1}^{N}\left(X_i-X_i'\right)\right\| &= \mathbb{E}\left\|\sum_{i=1}^{N}\varepsilon_i\left(X_i-X_i'\right)\right\| = \mathbb{E}\left\|\sum_{i=1}^{N}\varepsilon_i X_i-\sum_{i=1}^{N}\varepsilon_i X_i'\right\| \\ &\leq \mathbb{E}\left\|\sum_{i=1}^{N}\varepsilon_i X_i\right\|+\mathbb{E}\left\|\sum_{i=1}^{N}\varepsilon_i X_i'\right\| = 2\,\mathbb{E}\left\|\sum_{i=1}^{N}\varepsilon_i X_i\right\|.\end{aligned}$$

Lower bound. Conditionally on $\varepsilon=(\varepsilon_1,\ldots,\varepsilon_N)$, the vectors $\sum_{i=1}^{N}\varepsilon_i X_i$ and $-\sum_{i=1}^{N}\varepsilon_i X_i'$ are independent and the latter has mean zero, so Lemma 1.4.1 applies conditionally and gives, after averaging over ε,

$$\mathbb{E}\left\|\sum_{i=1}^{N}\varepsilon_i X_i\right\| \leq \mathbb{E}\left\|\sum_{i=1}^{N}\varepsilon_i X_i-\sum_{i=1}^{N}\varepsilon_i X_i'\right\| = \mathbb{E}\left\|\sum_{i=1}^{N}\varepsilon_i\left(X_i-X_i'\right)\right\|.$$

Applying (1.39) once more and then the triangle inequality,

$$\mathbb{E}\left\|\sum_{i=1}^{N}\varepsilon_i\left(X_i-X_i'\right)\right\| = \mathbb{E}\left\|\sum_{i=1}^{N}\left(X_i-X_i'\right)\right\| \leq \mathbb{E}\left\|\sum_{i=1}^{N}X_i\right\|+\mathbb{E}\left\|\sum_{i=1}^{N}X_i'\right\| = 2\,\mathbb{E}\left\|\sum_{i=1}^{N}X_i\right\|. \qquad \square$$

Remark 1.4.7. The zero-mean assumption is needed for both bounds: in the upper bound to subtract $\mathbb{E}\left[\sum_i X_i'\right]$, and in the lower bound to guarantee that $\mathbb{E}\left[\sum_i \varepsilon_i X_i' \mid \varepsilon\right]=0$.

1.5 Sub-exponential Random Variables

We now turn to the class of sub-exponential variables, which are defined by a slightly milder condition on the moment generating function.

Definition 1.5.1 ([1, Definition 2.7])**.** A random variable X with mean $\mu=\mathbb{E}[X]$ is *sub-exponential* if there are non-negative parameters (ν,α) such that

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\frac{\nu^2\lambda^2}{2}} \qquad \text{for all } |\lambda|<\frac{1}{\alpha}. \tag{1.40}$$

Remark 1.5.2. A σ-sub-Gaussian variable is $(\sigma,0)$-sub-exponential, where we interpret $1/0$ as $+\infty$.

Example 1.5.3 (Sub-exponential but not sub-Gaussian)**.** Let $Z\sim N(0,1)$ and consider $X=Z^2$, so that $\mathbb{E}[X]=1$. For $\lambda<\frac{1}{2}$,

$$\begin{aligned}\mathbb{E}\left[e^{\lambda(X-1)}\right] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{\lambda(z^2-1)}e^{-\frac{z^2}{2}}\,dz \\ &= \frac{e^{-\lambda}}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-\frac{(1-2\lambda)z^2}{2}}\,dz = \frac{e^{-\lambda}}{\sqrt{1-2\lambda}},\end{aligned} \tag{1.41}$$

where the last equality uses $\int_{-\infty}^{+\infty}e^{-az^2/2}dz=\sqrt{2\pi/a}$ with $a=1-2\lambda>0$. For $\lambda\geq\frac{1}{2}$ the moment generating function is infinite, so X is not sub-Gaussian. On the other hand,

$$\frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \leq e^{2\lambda^2} = e^{\frac{2^2\lambda^2}{2}} \qquad \text{where } |\lambda|<\tfrac{1}{4} \qquad \text{(check this!)}, \tag{1.42}$$

so that $\mathbb{E}\left[e^{\lambda(X-1)}\right]\leq e^{\frac{2^2\lambda^2}{2}}$ for all $|\lambda|<\frac{1}{4}$, i.e. X is $(2,4)$-sub-exponential.

Exercises

1. Verify the inequality (1.42), that is, show that $\frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \le e^{2\lambda^2}$ for all $|\lambda| < \frac{1}{4}$.

Proposition 1.5.4 (Sub-exponential tail bound)**.** *Suppose X is sub-exponential with parameters (ν, α) and mean μ. Then*

$$\mathbb{P}\left[X-\mu \ge t\right] \le \begin{cases} e^{-\frac{t^2}{2\nu^2}}, & \text{if } 0 \le t \le \frac{\nu^2}{\alpha}, \\ e^{-\frac{t}{2\alpha}}, & \text{for } t > \frac{\nu^2}{\alpha}, \end{cases} \quad = \quad e^{-\min\left\{\frac{t^2}{2\nu^2}, \frac{t}{2\alpha}\right\}}. \tag{1.43}$$

Proof. Assume $\mu = 0$ without loss of generality. By the Chernoff bound from Lecture 1 and Definition 1.5.1,

$$\begin{aligned} \log \mathbb{P}\left[X \ge t\right] &\le \inf_{0 \le \lambda < \frac{1}{\alpha}} \left\{\log \mathbb{E}\left[e^{\lambda X}\right] - \lambda t\right\} \le \inf_{0 \le \lambda < \frac{1}{\alpha}} \left\{\log e^{\frac{\nu^2\lambda^2}{2}} - \lambda t\right\} \\ &= \inf_{0 \le \lambda < \frac{1}{\alpha}} \underbrace{\left\{\frac{\nu^2\lambda^2}{2} - \lambda t\right\}}_{=:\, g(\lambda)}. \end{aligned} \tag{1.44}$$

The function g is a convex parabola (see Figure 1.2) with

$$g'(\lambda) = \nu^2\lambda - t = 0 \qquad \Longleftrightarrow \qquad \lambda = \frac{t}{\nu^2}.$$

① If $\frac{t}{\nu^2} < \frac{1}{\alpha}$ (i.e. $t < \frac{\nu^2}{\alpha}$), the unconstrained minimizer lies in the feasible region, so

$$\inf_{0 \le \lambda < \frac{1}{\alpha}} g(\lambda) = g\left(\frac{t}{\nu^2}\right) = \frac{\nu^2}{2} \cdot \frac{t^2}{\nu^4} - \frac{t}{\nu^2} \cdot t = -\frac{t^2}{2\nu^2}.$$

② If $\frac{t}{\nu^2} \ge \frac{1}{\alpha}$ (i.e. $t \ge \frac{\nu^2}{\alpha}$), then g is decreasing on $[0, \frac{t}{\nu^2})$ and the infimum is approached at $\lambda = \frac{1}{\alpha}$, giving

$$\inf_{0 \le \lambda < \frac{1}{\alpha}} g(\lambda) = g\left(\frac{1}{\alpha}\right) = \frac{\nu^2}{2} \cdot \frac{1}{\alpha^2} - \frac{1}{\alpha} t = \frac{\nu^2}{2\alpha^2} - \frac{t}{\alpha} \le -\frac{t}{2\alpha},$$

where the last step uses $\nu^2 < \alpha t$. ☐

1.5.1 Bernstein's Condition

We say that X satisfies *Bernstein's condition with parameter b* if

$$\left|\mathbb{E}\left[(X-\mu)^k\right]\right| \le \frac{1}{2} k!\, \mathrm{Var}(X)\, b^{k-2}, \qquad \text{for any } k = 2, 3, \dots, \tag{1.45}$$

where $\mu = \mathbb{E}[X]$.

Example 1.5.5. If (1.45) is satisfied, then X is sub-exponential. Indeed, writing $\sigma^2 = \mathrm{Var}(X)$ and using the Taylor expansion of the exponential together with $\mathbb{E}[X-\mu] = 0$,

$$\begin{aligned} \mathbb{E}\left[e^{\lambda(X-\mu)}\right] &= \mathbb{E}\left[\sum_{k=0}^{\infty} \frac{(\lambda(X-\mu))^k}{k!}\right] = 1 + 0 + \mathbb{E}\left[\sum_{k=2}^{\infty} \frac{\lambda^k (X-\mu)^k}{k!}\right] \\ &= 1 + \sum_{k=2}^{\infty} \frac{\lambda^k\, \mathbb{E}\left[(X-\mu)^k\right]}{k!} = 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty} \frac{\lambda^k\, \mathbb{E}\left[(X-\mu)^k\right]}{k!} \\ &\le 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty} \frac{|\lambda|^k \cdot \frac{1}{2} k!\, \sigma^2 b^{k-2}}{k!} = 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^2\sigma^2}{2} \sum_{k=3}^{\infty} \left(|\lambda|\, b\right)^{k-2} = 1 + \frac{\lambda^2\sigma^2}{2} \sum_{s=0}^{\infty} \left(|\lambda|\, b\right)^s. \end{aligned} \tag{1.46}$$

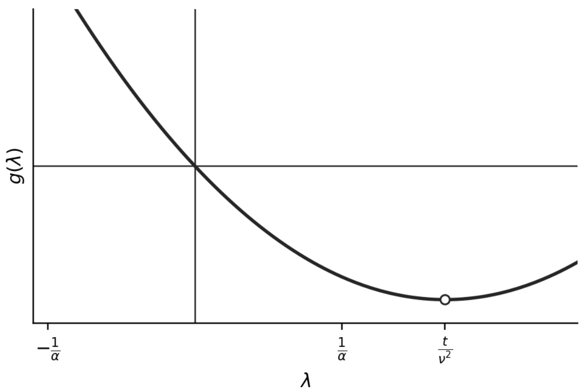

Figure 1.2: The function $g(\lambda) = \frac{\nu^2\lambda^2}{2} - \lambda t$ appearing in (1.44), in the case $\frac{t}{\nu^2} > \frac{1}{\alpha}$. The open circle marks the unconstrained minimizer $\frac{t}{\nu^2}$, which lies outside the feasible region $|\lambda| < \frac{1}{\alpha}$.

For $|\lambda| < \frac{1}{b}$ the geometric series sums to $\frac{1}{1-|\lambda|b}$, so that

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \le 1 + \frac{\lambda^2\sigma^2/2}{1-b\,|\lambda|} \le e^{\frac{\lambda^2\sigma^2/2}{1-b|\lambda|}}, \tag{1.47}$$

using $1 + t \le e^t$. Restricting further to $|\lambda| < \frac{1}{2b}$ gives $1 - b\,|\lambda| > \frac{1}{2}$, and therefore

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \le e^{\lambda^2\sigma^2} = e^{\frac{\lambda^2\left(\sqrt{2}\sigma\right)^2}{2}}, \qquad \text{for all } |\lambda| < \frac{1}{2b},$$

i.e. X is sub-exponential with parameters $(\nu, \alpha) = (\sqrt{2}\sigma, 2b)$.

Proposition 1.5.6 (Bernstein-type bound)**.** *Suppose X satisfies Bernstein's condition* (1.45) *with parameter b, mean μ, and variance σ^2. Then*

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \le \exp\left(\frac{\lambda^2\sigma^2/2}{1-b\,|\lambda|}\right) \qquad \textit{for all } |\lambda| < \frac{1}{b}, \tag{1.48}$$

and, moreover, the concentration inequality

$$\mathbb{P}\left(|X-\mu| > t\right) \le 2\exp\left(-\frac{t^2}{2(\sigma^2+bt)}\right) \qquad \textit{for all } t \ge 0. \tag{1.49}$$

Proof. Equation (1.48) directly follows from (1.46).

Remark 1.5.7. Restricting further to $|\lambda| < \frac{1}{2b}$ gives $b\,|\lambda| < \frac{1}{2}$, hence $1 - b\,|\lambda| > \frac{1}{2}$ and

$$\mathbb{E}\left[e^{\lambda X}\right] \le \exp\left(\frac{\lambda^2\sigma^2/2}{1/2}\right) = e^{\lambda^2\sigma^2} = e^{\frac{\lambda^2(\sqrt{2}\sigma)^2}{2}}, \qquad |\lambda| < \frac{1}{2b},$$

which recovers the Topic 3 conclusion that X is $(\sqrt{2}\sigma, 2b)$-sub-exponential. Bound (1.48) is sharper: it holds on the full range $|\lambda| < 1/b$ instead of only $|\lambda| < 1/(2b)$, and this extra room is exactly what lets us optimize the Chernoff bound below.

Proof of (1.49). By the Chernoff bound and (1.48), for any $t \geq 0$,

$$\log \mathbb{P}\left(X-\mu \geq t\right) \leq \inf_{0 \leq \lambda < 1/b} \left\{ \log \mathbb{E}\left[e^{\lambda(X-\mu)}\right] - \lambda t \right\} \leq \inf_{0 \leq \lambda < 1/b} \underbrace{\left\{ \frac{\lambda^2 \sigma^2}{2(1-b\lambda)} - \lambda t \right\}}_{=:\, \theta(\lambda)}. \tag{1.50}$$

Rather than optimizing θ exactly, it suffices to evaluate it at the specific point

$$\lambda^* := \frac{t}{\sigma^2 + bt}.$$

Since $t, \sigma^2, b \geq 0$ we have $\lambda^* \geq 0$, and $\lambda^* < 1/b \iff bt < \sigma^2 + bt \iff \sigma^2 > 0$, which holds whenever X is non-degenerate (if $\sigma^2 = 0$ then $X \equiv \mu$ a.s. and (1.49) is trivial). At $\lambda = \lambda^*$,

$$1 - b\lambda^* = 1 - \frac{bt}{\sigma^2 + bt} = \frac{\sigma^2}{\sigma^2 + bt},$$

so that

$$\theta(\lambda^*) = \frac{\sigma^2 (\lambda^*)^2 (\sigma^2 + bt)}{2\sigma^2} - \lambda^* t = \frac{(\lambda^*)^2(\sigma^2 + bt)}{2} - \lambda^* t = \frac{t^2}{2(\sigma^2 + bt)} - \frac{t^2}{\sigma^2 + bt} = -\frac{t^2}{2(\sigma^2 + bt)}.$$

Combining with (1.50) (using that λ^* is feasible),

$$\mathbb{P}\left(X - \mu \geq t\right) \leq \exp\left(-\frac{t^2}{2(\sigma^2 + bt)}\right), \qquad t \geq 0. \tag{1.51}$$

Finally, $-X$ also satisfies Bernstein's condition with the same (σ^2, b), since $\left|\mathbb{E}[(-(X-\mu))^k]\right| = \left|\mathbb{E}[(X-\mu)^k]\right|$ and $\mathrm{Var}(-X) = \sigma^2$. Applying (1.51) to $-X$ gives $\mathbb{P}(X - \mu \leq -t) \leq \exp(-t^2/(2(\sigma^2 + bt)))$ as well, and adding the two one-sided bounds yields (1.49). $\square$

Exercise 1.5.8. Compare the bound (1.49) with the bound one would obtain by plugging the (weaker) parameters $(\sqrt{2}\sigma, 2b)$ derived in the Remark above directly into the general sub-exponential tail bound of Topic 3. Show that (1.49) is never worse, and can be substantially tighter when $bt \ll \sigma^2$.

1.5.2 Sums of Independent Sub-exponential Random Variables

Recall that X with mean μ is *sub-exponential with parameters* (ν, α) if $\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\nu^2\lambda^2/2}$ for all $|\lambda| < 1/\alpha$, and that in this case

$$\mathbb{P}\left(|X - \mu| > t\right) \leq \begin{cases} 2e^{-t^2/(2\nu^2)}, & 0 \leq t \leq \nu^2/\alpha, \\ 2e^{-t/(2\alpha)}, & t > \nu^2/\alpha. \end{cases} \tag{1.52}$$

Lemma 1.5.9. *Let $X_1, \ldots, X_n$ be independent random variables, with X_i sub-exponential with parameters (ν_i, d_i). Then $S_n := \sum_{i=1}^n X_i$ is sub-exponential with parameters*

$$\left(\sqrt{\sum_{i=1}^n \nu_i^2},\ \max_{1 \leq i \leq n} d_i\right).$$

and

$$\mathbb{P}\left(|S_n - \mathbb{E}[S_n]| > t\right) \leq \begin{cases} 2e^{-t^2/(2V^2)}, & 0 \leq t \leq V^2/d, \\ 2e^{-t/(2d)}, & t > V^2/d, \end{cases} \qquad V^2 = \sum_{i=1}^n \nu_i^2, \quad d = \max_{1 \leq i \leq n} d_i. \tag{1.53}$$

Proof. Write $d := \max_{1\le i\le n} d_i$ and $V^2 := \sum_{i=1}^n \nu_i^2$. For any $|\lambda| < 1/d = \min_{1\le i\le n}(1/d_i)$, we have $|\lambda| < 1/d_i$ for every i, so by independence,

$$\mathbb{E}\left[e^{\lambda(S_n-\mathbb{E}[S_n])}\right] = \mathbb{E}\left[e^{\lambda\sum_{i=1}^n (X_i-\mathbb{E}[X_i])}\right] = \prod_{i=1}^n \mathbb{E}\left[e^{\lambda(X_i-\mathbb{E}[X_i])}\right] \le \prod_{i=1}^n e^{\nu_i^2\lambda^2/2} = e^{\lambda^2V^2/2}.$$

Hence S_n is (V,d)-sub-exponential.

By (1.52) applied with $(\nu,\alpha)=(V,d)$, we obtain the tail bound

$$\mathbb{P}\left(|S_n-\mathbb{E}[S_n]|>t\right) \le \begin{cases} 2e^{-t^2/(2V^2)}, & 0\le t\le V^2/d, \\ 2e^{-t/(2d)}, & t > V^2/d, \end{cases} \qquad V^2=\sum_{i=1}^n \nu_i^2, \quad d=\max_{1\le i\le n} d_i.$$

□

Example 1.5.10 (Concentration of χ^2 variables)**.** Let $Z_1,\dots,Z_n$i.i.d. $N(0,1)$ and let $Y := \sum_{i=1}^n Z_i^2 \sim \chi_n^2$, so that $\mathbb{E}[Y]=n$. Recall from Topic 3 that each Z_i^2 has mean 1 and is $(2,4)$-sub-exponential, i.e. $\mathbb{E}[e^{\lambda(Z_i^2-1)}] \le e^{2\lambda^2}$ for all $|\lambda|<1/4$. Applying Lemma 1.5.9 with $\nu_i=2$, $d_i=4$ for every i, Y is $(\sqrt{4n},4)=(2\sqrt{n},4)$-sub-exponential; here $V^2=4n$ and $d=4$, so the threshold in (1.53) is $V^2/d=n$. Thus

$$\mathbb{P}\left(|Y-n|\ge s\right) \le \begin{cases} 2e^{-s^2/(8n)}, & 0\le s\le n, \\ 2e^{-s/8}, & s>n. \end{cases}$$

Substituting $s=nt$ (so $s\le n \iff t\le 1$) gives

$$\mathbb{P}\left(\left|\frac{Y}{n}-1\right|\ge t\right) \le \begin{cases} 2e^{-nt^2/8}, & 0\le t\le 1, \\ 2e^{-nt/8}, & t>1. \end{cases} \tag{1.54}$$

1.5.3 Application: The Johnson–Lindenstrauss Lemma

Fix distinct points $u^1,\dots,u^N\in\mathbb{R}^d$, where d is large. We would like a map $F:\mathbb{R}^d\to\mathbb{R}^m$ with $m\ll d$, so that we only need to store the compressed points $\{F(u^1),\dots,F(u^N)\}$, while approximately preserving all pairwise (squared) distances: given a distortion parameter $\delta\in(0,1)$, we want

$$1-\delta \;\le\; \frac{\left\|F(u^i)-F(u^j)\right\|_2^2}{\left\|u^i-u^j\right\|_2^2} \;\le\; 1+\delta, \qquad \text{for all } i\ne j. \tag{1.55}$$

Construction. Let $X\in\mathbb{R}^{m\times d}$ have i.i.d. $N(0,1)$ entries, and write its rows as $X_1,\dots,X_m\in\mathbb{R}^d$, so that $X_1,\dots,X_m$i.i.d. $N(0,I_d)$. Define the (random) linear map

$$F(u) := \frac{Xu}{\sqrt{m}}, \qquad u\in\mathbb{R}^d.$$

Step 1: distribution of $\|Xu\|_2^2$. Fix $u\in\mathbb{R}^d$, $u\ne 0$. Writing $Xu=(\langle X_1,u\rangle,\dots,\langle X_m,u\rangle)^\top$, for each i and any unit vector $v=u/\|u\|_2$,

$$\langle X_i,v\rangle = \sum_{k=1}^d X_{i,k}v_k \sim N\left(0,\sum_{k=1}^d v_k^2\right) = N(0,1) \qquad (\text{check: } \|v\|_2=1), \tag{1.56}$$

since $X_i\sim N(0,I_d)$ and any linear combination of independent standard normals with coefficient vector of unit ℓ_2-norm is again standard normal. Moreover $\langle X_1,v\rangle,\dots,\langle X_m,v\rangle$ are independent because $X_1,\dots,X_m$ are independent. Hence

$$Y := \frac{\|Xu\|_2^2}{\|u\|_2^2} = \sum_{i=1}^m \frac{\langle X_i,u\rangle^2}{\|u\|_2^2} = \sum_{i=1}^m \langle X_i,v\rangle^2 \sim \chi_m^2. \tag{1.57}$$

Step 2: concentration for a single vector. Since $\|F(u)\|_2^2 / \|u\|_2^2 = \|Xu\|_2^2 / (m\,\|u\|_2^2) = Y/m$, applying (1.54) (with n replaced by m) gives, for any $\delta \in (0,1)$,

$$\mathbb{P}\left(\frac{\|F(u)\|_2^2}{\|u\|_2^2} \notin (1-\delta, 1+\delta)\right) = \mathbb{P}\left(\left|\frac{Y}{m} - 1\right| \geq \delta\right) \leq 2e^{-m\delta^2/8}. \tag{1.58}$$

Step 3: union bound over all pairs. Apply (1.58) with $u = u^i - u^j$ for each of the $\binom{N}{2}$ unordered pairs $i \neq j$ (assuming $u^i \neq u^j$):

$$\mathbb{P}\left(\frac{\left\|F(u^i) - F(u^j)\right\|_2^2}{\|u^i - u^j\|_2^2} \notin (1-\delta, 1+\delta)\right) \leq 2e^{-m\delta^2/8}.$$

Let

$$I := \bigcup_{i<j}\left\{\frac{\left\|F(u^i) - F(u^j)\right\|_2^2}{\|u^i - u^j\|_2^2} \notin (1-\delta, 1+\delta)\right\}$$

be the event that (1.55) fails for at least one pair. By the union bound,

$$\mathbb{P}(I) \leq \sum_{i<j} \mathbb{P}\left(\frac{\left\|F(u^i) - F(u^j)\right\|_2^2}{\|u^i - u^j\|_2^2} \notin (1-\delta, 1+\delta)\right) \leq \binom{N}{2} \cdot 2e^{-m\delta^2/8} = N(N-1)\, e^{-m\delta^2/8}. \tag{1.59}$$

Step 4: choosing m. Given a target failure probability $\varepsilon \in (0,1)$, we want $\mathbb{P}(I) < \varepsilon$. By (1.59), it suffices that

$$N(N-1)e^{-m\delta^2/8} < \varepsilon \quad \Longleftrightarrow \quad m > \frac{8}{\delta^2} \log \frac{N(N-1)}{\varepsilon}.$$

Since $N(N-1) \leq N^2$, a cleaner sufficient condition is

$$m > \frac{16}{\delta^2} \log \frac{N}{\varepsilon}. \tag{1.60}$$

Theorem 1.5.11 (Johnson–Lindenstrauss Lemma)**.** *Let distinct $u^1, \ldots, u^N \in \mathbb{R}^d$ and let $\delta, \varepsilon \in (0,1)$. If*

$$m > \frac{16}{\delta^2} \log \frac{N}{\varepsilon},$$

then, with $X \in \mathbb{R}^{m \times d}$ having i.i.d. $N(0,1)$ entries and $F(u) := Xu/\sqrt{m}$, with probability at least $1-\varepsilon$ we have

$$1 - \delta \leq \frac{\left\|F(u^i) - F(u^j)\right\|_2^2}{\|u^i - u^j\|_2^2} \leq 1 + \delta \qquad \textit{simultaneously for all } i \neq j.$$

Proof. Immediate from (1.59) and (1.60): the choice of m forces $\mathbb{P}(I) < \varepsilon$, i.e. (1.55) holds for every pair on the complementary event, which has probability at least $1-\varepsilon$. $\square$

Bibliography

[1] Martin J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Cambridge University Press, 2019.

[2] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*, Cambridge University Press, 2018.