# 2

# Concentration of Sums of Independent Random Variables

This chapter introduces the reader to the rich topic of concentration inequalities. After motivating the subject in Section 2.1, we prove some basic concentration inequalities: Hoeffding's in Sections 2.2 and 2.6, Chernoff's in Section 2.3, and Bernstein's in Section 2.8. Another goal of this chapter is to introduce two important classes of distributions: sub-gaussian in Section 2.5 and sub-exponential in Section 2.7. These classes form a natural "habitat" in which many results of high-dimensional probability and its applications will be developed. We give two quick applications of concentration inequalities for randomized algorithms in Section 2.2 and random graphs in Section 2.4. Many more applications are given later in the book.

## 2.1 Why Concentration Inequalities?

Concentration inequalities quantify how a random variable $X$ deviates around its mean $\mu$. They usually take the form of two-sided bounds for the tails of $X - \mu$, such as

$$\mathbb{P}\left\{|X-\mu| > t\right\} \le \text{something small.}$$

The simplest concentration inequality is Chebyshev's inequality (Corollary 1.2.5). It is very general but often too weak. Let us illustrate this with the example of the binomial distribution.

**Question 2.1.1** *Toss a fair coin $N$ times. What is the probability that we get at least* $3N/4$ *heads?*

Let $S_N$ denote the number of heads. Then

$$\mathbb{E}\, S_N = \frac{N}{2}, \quad \operatorname{Var}(S_N) = \frac{N}{4}.$$

Chebyshev's inequality bounds the probability of getting at least $3N/4$ heads as follows:

$$\mathbb{P}\left\{S_N \ge \frac{3}{4}N\right\} \le \mathbb{P}\left\{\left|S_N - \frac{N}{2}\right| \ge \frac{N}{4}\right\} \le \frac{4}{N}. \tag{2.1}$$

So the probability converges to zero at least *linearly* in $N$.

Is this the right rate of decay, or we should expect something faster? Let us approach the same question using the central limit theorem. To do this, we represent $S_N$ as a sum of independent random variables:

$$S_N = \sum_{i=1}^{N} X_i$$

where the $X_i$ are independent Bernoulli random variables with parameter $1/2$, i.e. $\mathbb{P}\left\{X_i = 0\right\} = \mathbb{P}\left\{X_i = 1\right\} = 1/2$. (These $X_i$ are the indicators of heads.) The De Moivre–Laplace central limit theorem (1.7) states that the distribution of the normalized number of heads,

$$Z_N = \frac{S_N - N/2}{\sqrt{N/4}},$$

converges to the standard normal distribution $N(0, 1)$. Thus we should anticipate that for large $N$, we have

$$\mathbb{P}\left\{S_N \geq 3N/4\right\} = \mathbb{P}\left\{Z_N \geq \sqrt{N/4}\right\} \approx \mathbb{P}\left\{g \geq \sqrt{N/4}\right\} \tag{2.2}$$

where $g \sim N(0, 1)$. To understand how this quantity decays in $N$, we will now obtain a good bound on the tails of the normal distribution.

**Proposition 2.1.2** (Tails of the normal distribution) *Let $g \sim N(0, 1)$. Then, for all $t > 0$, we have*

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

*In particular, for $t \geq 1$ the tail is bounded by the density:*

$$\mathbb{P}\left\{g \geq t\right\} \leq \frac{1}{\sqrt{2\pi}} e^{-t^2/2}. \tag{2.3}$$

*Proof* To obtain an upper bound on the tail

$$\mathbb{P}\left\{g \geq t\right\} = \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-x^2/2}\, dx,$$

let us make the change of variables $x = t + y$. This gives

$$\mathbb{P}\left\{g \geq t\right\} = \frac{1}{\sqrt{2\pi}} \int_0^\infty e^{-t^2/2}\, e^{-ty}\, e^{-y^2/2}\, dy \leq \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \int_0^\infty e^{-ty} dy,$$

where we have used that $e^{-y^2/2} \leq 1$. Since the last integral equals $1/t$, the desired upper bound on the tail follows.

The lower bound follows from the identity

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

This completes the proof. ■

Returning to (2.2), we see that we should expect the probability of having at least $3N/4$ heads to be smaller than

$$\frac{1}{\sqrt{2\pi}} e^{-N/8}. \tag{2.4}$$

This quantity decays to zero *exponentially* fast in $N$, which is much better than the linear decay in (2.1) that follows from Chebyshev's inequality.

Unfortunately, (2.4) does not follow rigorously from the central limit theorem. Although the approximation by the normal density in (2.2) is valid, the error of approximation cannot be ignored. And, unfortunately, *the error decays too slowly* – even more slowly than linearly in $N$. This can be seen from the following sharp quantitative version of the central limit theorem.

**Theorem 2.1.3** (Berry–Esseen central limit theorem) *In the setting of Theorem 1.3.2, for every $N$ and every $t \in \mathbb{R}$ we have*

$$\left|\mathbb{P}\left\{Z_N \geq t\right\} - \mathbb{P}\left\{g \geq t\right\}\right| \leq \frac{\rho}{\sqrt{N}}.$$

*Here $\rho = \mathbb{E}\,|X_1 - \mu|^3/\sigma^3$ and $g \sim N(0, 1)$.*

Thus the approximation error in (2.2) is of order $1/\sqrt{N}$, which ruins the desired exponential decay (2.4).

Can we improve the approximation error involved in using the central limit theorem? In general, no. If $N$ is even then the probability of getting exactly $N/2$ heads is

$$\mathbb{P}\left\{S_N = N/2\right\} = 2^{-N}\binom{N}{N/2} \sim \frac{1}{\sqrt{N}};$$

the last estimate can be obtained using Stirling's approximation. (Do it!) Hence, $\mathbb{P}\left\{Z_N = 0\right\} \sim 1/\sqrt{N}$. On the other hand, since the normal distribution is continuous, we have $\mathbb{P}\left\{g = 0\right\} = 0$. Thus the approximation error here has to be of order $1/\sqrt{N}$.

Let us summarize our situation. The central limit theorem offers an approximation of a sum of independent random variables $S_N = X_1 + \cdots + X_N$ by the normal distribution. The normal distribution is especially nice owing to its very light, exponentially decaying, tails. At the same time, the error of approximation in the central limit theorem decays too slowly, even more slowly than linear. This large error is a roadblock toward proving concentration properties for random variables $S_N$ with light, exponentially decaying, tails.

In order to resolve this issue, we will develop alternative, direct, approaches to concentration which bypass the central limit theorem.

**Exercise 2.1.4** (Truncated normal distribution)☕ Let $g \sim N(0, 1)$. Show that, for all $t \geq 1$, we have

$$\mathbb{E}\, g^2 \mathbf{1}_{\{g>t\}} = t\,\frac{1}{\sqrt{2\pi}}e^{-t^2/2} + \mathbb{P}\left\{g > t\right\} \leq \left(t + \frac{1}{t}\right)\frac{1}{\sqrt{2\pi}}e^{-t^2/2}.$$ ☞

## 2.2 Hoeffding's Inequality

We start with a particularly simple concentration inequality, which holds for sums of i.i.d. *symmetric Bernoulli* random variables.

**Definition 2.2.1** (Symmetric Bernoulli distribution) A random variable $X$ has a *symmetric Bernoulli* distribution (also called a *Rademacher* distribution) if it takes values $-1$ and $1$ with probabilities $1/2$ each, i.e.,

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

Clearly, a random variable $X$ has the (usual) Bernoulli distribution with parameter $1/2$ if and only if $Z=2X-1$ has a symmetric Bernoulli distribution.

**Theorem 2.2.2** (Hoeffding's inequality) *Let $X_1,\dots,X_N$ be independent symmetric Bernoulli random variables, and let $a=(a_1,\dots,a_N)\in\mathbb{R}^N$. Then, for any $t\ge 0$, we have*

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

*Proof* We can assume without loss of generality that $\|a\|_2=1$. (Why?)

Let us recall how we deduced Chebyshev's inequality (Corollary 1.2.5): we squared both sides and applied Markov's inequality. Let us do something similar here. But instead of squaring both sides, let us multiply by a fixed parameter $\lambda>0$ (to be chosen later) and exponentiate. This gives

$$\begin{aligned}\mathbb{P}\left\{\sum_{i=1}^N a_i X_i \ge t\right\} &= \mathbb{P}\left\{\exp\Big(\lambda\sum_{i=1}^N a_i X_i\Big)\ge \exp(\lambda t)\right\}\\ &\le e^{-\lambda t}\,\mathbb{E}\exp\Big(\lambda\sum_{i=1}^N a_i X_i\Big). \qquad (2.5)\end{aligned}$$

In the last step we applied Markov's inequality (Proposition 1.2.4).

We have thus reduced the problem to bounding the *moment generating function* (MGF) of the sum $\sum_{i=1}^N a_i X_i$. Recall that the MGF of the sum is the product of the MGFs of the terms; this follows immediately from the independence of the $X_i$. Thus

$$\mathbb{E}\exp\Big(\lambda\sum_{i=1}^N a_i X_i\Big)=\prod_{i=1}^N \mathbb{E}\exp(\lambda a_i X_i). \tag{2.6}$$

Let us fix $i$. Since $X_i$ takes values $-1$ and $1$ with probabilities $1/2$ each, we have

$$\mathbb{E}\exp(\lambda a_i X_i)=\frac{\exp(\lambda a_i)+\exp(-\lambda a_i)}{2}=\cosh(\lambda a_i).$$

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

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

☞

This bound shows that

$$\mathbb{E}\exp(\lambda a_i X_i)\le\exp(\lambda^2 a_i^2/2).$$

Substituting into (2.6) and then into (2.5), we obtain

$$\begin{aligned}\mathbb{P}\left\{\sum_{i=1}^{N} a_i X_i \geq t\right\} \leq e^{-\lambda t} \prod_{i=1}^{N} \exp(\lambda^2 a_i^2/2) &= \exp\Big(-\lambda t + \frac{\lambda^2}{2}\sum_{i=1}^{N} a_i^2\Big)\\ &= \exp\Big(-\lambda t + \frac{\lambda^2}{2}\Big).\end{aligned}$$

In the last identity, we used the assumption that $\|a\|_2 = 1$.

This bound holds for arbitrary $\lambda > 0$. It remains to optimize in $\lambda$; the minimum is clearly attained for $\lambda = t$. With this choice, we obtain

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

This completes the proof of Hoeffding's inequality. ■

We can view Hoeffding's inequality as a concentration version of the central limit theorem. Indeed, the most that we may expect from a concentration inequality is that the tail of $\sum a_i X_i$ behaves similarly to the tail of the normal distribution. And, for all practical purposes, Hoeffding's tail bound does that. With the normalization $\|a\|_2 = 1$, Hoeffding's inequality provides the tail $e^{-t^2/2}$, which is exactly the same as the bound for the standard normal tail in (2.3). 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.

Armed with Hoeffding's inequality, we can now return to Question 2.1.1 regarding the bounding of the probability of at least $3N/4$ heads in $N$ tosses of a fair coin. After rescaling from Bernoulli to symmetric Bernoulli, we obtain that this probability is *exponentially small* in $N$, namely

$$\mathbb{P}\left\{\text{at least } 3N/4 \text{ heads}\right\} \leq \exp(-N/8).$$

(Check this.)

**Remark 2.2.4** (Non-asymptotic results) 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 the value of $N$, the stronger the inequality becomes. As we will see later, the non-asymptotic nature of concentration inequalities like that of Hoeffding makes them attractive in applications in the data sciences, where $N$ often corresponds to the *sample size*.

We can easily derive a version of Hoeffding's inequality for *two-sided tails* $\mathbb{P}\left\{|S| \geq t\right\}$ where $S = \sum_{i=1}^{N} a_i X_i$. Indeed, applying Hoeffding's inequality for $-X_i$ instead of $X_i$, we obtain a bound on $\mathbb{P}\left\{-S \geq t\right\}$. Combining the two bounds, we obtain the bound

$$\mathbb{P}\left\{|S| \geq t\right\} = \mathbb{P}\left\{S \geq t\right\} + \mathbb{P}\left\{-S \geq t\right\}.$$

Thus the bound doubles, and we obtain:

**Theorem 2.2.5** (Hoeffding's inequality, two-sided) *Let $X_1, \dots, X_N$ be independent symmetric Bernoulli random variables, and let $a = (a_1, \dots, a_N) \in \mathbb{R}^N$. Then, for any $t > 0$, we have*

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

Our proof above of Hoeffding's inequality, which is based on bounding the moment generating function, is quite flexible. It applies far beyond the canonical example of the symmetric Bernoulli distribution. For example, the following extension of Hoeffding's inequality is valid for general bounded random variables.

**Theorem 2.2.6** (Hoeffding's inequality for general bounded random variables) *Let $X_1, \dots, X_N$ be independent random variables. Assume that $X_i \in [m_i, M_i]$ for every $i$. Then, for any $t > 0$, we have*

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

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

**Exercise 2.2.8** (Boosting randomized algorithms)☕☕ Imagine we have an algorithm for solving some decision problem (e.g., is a given number $p$ a prime?). Suppose that the algorithm makes a decision at random and returns the correct answer with probability $1/2 + \delta$, for some $\delta > 0$, which is just a bit better than a random guess. To improve the performance, we run the algorithm $N$ times and take the majority vote. Show that, for any $\varepsilon \in (0, 1)$, the answer is correct with probability $1 - \varepsilon$, as long as $N \ge (1/2)\delta^{-2} \ln(\varepsilon^{-1})$. ☞

**Exercise 2.2.9** (Robust estimation of the mean)☕☕☕ Suppose that we want to estimate the mean $\mu$ of a random variable $X$ from a sample $X_1, \dots, X_N$ drawn independently from the distribution of $X$. We want an $\varepsilon$-accurate estimate, i.e. one that falls in the interval $(\mu - \varepsilon, \mu + \varepsilon)$.

(a) Show that a sample of size $N = O(\sigma^2/\varepsilon^2)$ is sufficient to compute an $\varepsilon$-accurate estimate with probability at least 3/4, where $\sigma^2 = \operatorname{Var} X$.[1] ☞
(b) Show that a sample of size $N = O(\log(\delta^{-1})\sigma^2/\varepsilon^2)$ is sufficient to compute an $\varepsilon$-accurate estimate with probability at least $1 - \delta$. ☞

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

[1] More accurately, this claim means that there exists an absolute constant $C$ such that if $N \ge C\sigma^2/\varepsilon^2$ then $\mathbb{P}\left\{ |\hat{\mu} - \mu| \le \varepsilon \right\} \ge 3/4$. Here $\hat{\mu}$ is the sample mean; see the hint.

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

$$\mathbb{E}\exp(-tX_i) \le \frac{1}{t} \quad \text{for all } t > 0.$$

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

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

## 2.3 Chernoff's Inequality

As we have noted, Hoeffding's inequality is quite sharp for symmetric Bernoulli random variables. But the general form of Hoeffding's inequality (Theorem 2.2.6) is sometimes too conservative and does not give sharp results. This happens, for example, when the $X_i$ are Bernoulli random variables with parameters $p_i$ so small that we expect $S_N$ to have an approximately Poisson distribution according to Theorem 1.3.4. However, Hoeffding's inequality is not sensitive to the magnitudes of the $p_i$, and the Gaussian tail bound that it gives is very far from the true, Poisson, tail. In this section we study Chernoff's inequality, which is sensitive to the magnitudes of the $p_i$.

**Theorem 2.3.1** (Chernoff's inequality) *Let $X_i$ be independent Bernoulli random variables with parameters $p_i$. Consider their sum $S_N = \sum_{i=1}^N X_i$ and denote its mean by $\mu = \mathbb{E}\, S_N$. Then, for any $t > \mu$, we have*

$$\mathbb{P}\left\{S_N \ge t\right\} \le e^{-\mu}\left(\frac{e\mu}{t}\right)^t.$$

*Proof* We will use the same method – based on the moment generating function – as we did in the proof of Hoeffding's inequality, Theorem 2.2.2. We repeat the first steps of that argument, leading to (2.5) and (2.6): multiply both sides of the inequality $S_N \ge t$ by a parameter $\lambda$, exponentiate, and then use Markov's inequality and independence. This gives

$$\mathbb{P}\left\{S_N \ge t\right\} \le e^{-\lambda t} \prod_{i=1}^{N} \mathbb{E}\exp(\lambda X_i). \tag{2.7}$$

It remains to bound the MGF of each Bernoulli random variable $X_i$ separately. Since $X_i$ takes the value 1 with probability $p_i$ and the value 0 with probability $1 - p_i$, we have

$$\mathbb{E}\exp(\lambda X_i) = e^\lambda p_i + (1 - p_i) = 1 + (e^\lambda - 1)p_i \le \exp\left((e^\lambda - 1)p_i\right).$$

In the last step, we used the numeric inequality $1 + x \le e^x$. Consequently,

$$\prod_{i=1}^{N} \mathbb{E}\exp(\lambda X_i) \le \exp\left((e^\lambda - 1)\sum_{i=1}^{N} p_i\right) = \exp\left((e^\lambda - 1)\mu\right).$$

Substituting this into (2.7), we obtain

$$\mathbb{P}\left\{S_N \ge t\right\} \le e^{-\lambda t}\exp\left((e^\lambda - 1)\mu\right).$$

This bound holds for any $\lambda > 0$. Substituting the value $\lambda = \ln(t/\mu)$, which is positive by the assumption $t > \mu$, and simplifying the expression, we complete the proof. ■

**Exercise 2.3.2** (Chernoff's inequality: lower tails)☕☕ Modify the proof of Theorem 2.3.1 to obtain the following bound on the lower tail. For any $t < \mu$, we have

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

**Exercise 2.3.3** (Poisson tails)☕☕ Let $X \sim \text{Pois}(\lambda)$. Show that, for any $t > \lambda$, we have

$$\mathbb{P}\left\{X \geq t\right\} \leq e^{-\lambda}\left(\frac{e\lambda}{t}\right)^t. \tag{2.8}$$

☞

**Remark 2.3.4** (Poisson tails) Note that the Poisson tail bound (2.8) is quite sharp. Indeed, the probability mass function (1.8) of $X \sim \text{Pois}(\lambda)$ can be approximated via Stirling's formula $k! \sim \sqrt{2\pi k}(k/e)^k$ as follows:

$$\mathbb{P}\left\{X = k\right\} \sim \frac{1}{\sqrt{2\pi k}}\, e^{-\lambda}\left(\frac{e\lambda}{k}\right)^k. \tag{2.9}$$

So our bound (2.8) on the *entire tail* of $X$ has essentially the same form as the probability of hitting *one value* $k$ (the smallest value) in that tail. The difference between these two quantities is the factor $\sqrt{2\pi k}$, which is negligible since both these quantities are exponentially small in $k$.

**Exercise 2.3.5** (Chernoff's inequality: small deviations)☕☕☕ Show that, in the setting of Theorem 2.3.1, for $\delta \in (0, 1]$ we have

$$\mathbb{P}\left\{|S_N - \mu| \geq \delta\mu\right\} \leq 2e^{-c\mu\delta^2},$$

where $c > 0$ is an absolute constant. ☞

**Exercise 2.3.6** (Poisson distribution near the mean)☕ Let $X \sim \text{Pois}(\lambda)$. Show that for $t \in (0, \lambda]$, we have

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

☞

**Remark 2.3.7** (Large and small deviations) Exercises 2.3.3 and 2.3.6 indicate two different behaviors of the tail of the Poisson distribution $\text{Pois}(\lambda)$. In the small-deviation regime, near the mean $\lambda$, the tail of $\text{Pois}(\lambda)$ is like that for the normal distribution $N(\lambda, \lambda)$. In the large-deviation regime, far to the right from the mean, the tail is heavier and decays like $(\lambda/t)^t$; see Figure 2.1.

**Exercise 2.3.8** (Normal approximation to Poisson)☕☕ Let $X \sim \text{Pois}(\lambda)$. Show that, as $\lambda \to \infty$, we have

$$\frac{X - \lambda}{\sqrt{\lambda}} \to N(0, 1) \quad \text{in distribution.}$$

☞

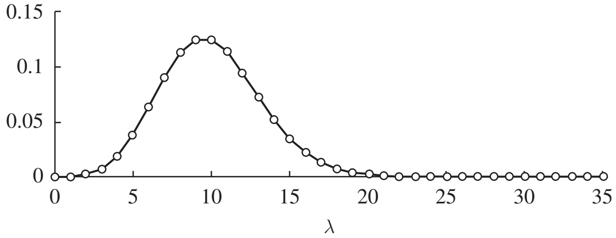

**Figure 2.1** The probability mass function of the Poisson distribution Pois($\lambda$) with $\lambda = 10$. The distribution is approximately normal near the mean $\lambda$, but to the right of the mean the tail is heavier.

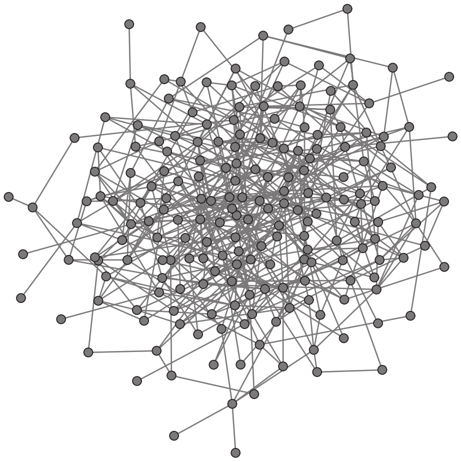

**Figure 2.2** A random graph from the Erdös–Rényi model $G(n, p)$ with $n = 200$ and $p = 1/40$.

## 2.4 Application: Degrees of Random Graphs

We now give an application of Chernoff's inequality to a classical object in probability: *random graphs*.

The most thoroughly studied model of random graphs is the classical *Erdös–Rényi model* $G(n, p)$, which is constructed on a set of $n$ vertices by connecting every pair of distinct vertices independently with probability $p$. Figure 2.2 shows an example of a random graph $G \sim G(n, p)$. In applications, the Erdös–Rényi model often appears as the simplest stochastic model for large, real-world, *networks*.

The *degree* of a vertex in the graph is the number of edges incident to that vertex. The expected degree of every vertex in $G(n, p)$ clearly equals

$$(n - 1)p =: d.$$

(Check!) We will show that relatively *dense graphs*, those where $d \gtrsim \log n$, are almost *regular*, with high probability, which means that the degrees of all vertices approximately equal $d$.

**Proposition 2.4.1** (Dense graphs are almost regular) *There is an absolute constant $C$ such that the following holds. Consider a random graph $G \sim G(n, p)$ with expected degree satisfying $d \geq C \log n$. Then, with high probability (for example,* 0.9*), the following occurs: all vertices of $G$ have degrees between* $0.9d$ *and* $1.1d$.

*Proof* The argument uses a combination of Chernoff's inequality and a *union bound*. Let us fix a vertex $i$ of the graph. The degree of $i$, which we denote $d_i$, is a sum of $n - 1$ independent $\mathrm{Ber}(p)$ random variables (the indicators of the edges incident to $i$). Thus we can apply Chernoff's inequality, which yields

$$\mathbb{P}\left\{|d_i - d| \geq 0.1d\right\} \leq 2e^{-cd}.$$

(Here, we have used the version of Chernoff's inequality given in Exercise 2.3.5.)

This bound holds for each fixed vertex $i$. Next, we can "unfix" $i$ by taking the union bound over all $n$ vertices. We obtain

$$\mathbb{P}\left\{\exists i \leq n\colon |d_i - d| \geq 0.1d\right\} \leq \sum_{i=1}^{n} \mathbb{P}\left\{|d_i - d| \geq 0.1d\right\} \leq n\, 2e^{-cd}.$$

If $d \geq C \log n$ for a sufficiently large absolute constant $C$, the probability is bounded by $0.1$. This means that, with probability 0.9, the complementary event occurs and we have

$$\mathbb{P}\left\{\forall i \leq n\colon |d_i - d| < 0.1d\right\} \geq 0.9.$$

This completes the proof. ■

Sparser graphs, those for which $d = o(\log n)$, are no longer almost regular, but there are still useful bounds on their degrees. The following series of exercises makes these claims clear. In all of them we shall assume that the graph size $n$ grows to infinity but we will not assume the connection probability $p$ to be constant in $n$.

**Exercise 2.4.2** (Bounding the degrees of sparse graphs)☕ Consider a random graph $G \sim G(n, p)$ with expected degrees $d = O(\log n)$. Show that with high probability (say, 0.9), all the vertices of $G$ have degree $O(\log n)$. ☞

**Exercise 2.4.3** (Bounding the degrees of very sparse graphs)☕☕ Consider a random graph $G \sim G(n, p)$ with expected degrees $d = O(1)$. Show that with high probability (say, 0.9), all the vertices of $G$ have degree

$$O\Big(\frac{\log n}{\log \log n}\Big).$$

Now we pass to the lower bounds. The next exercise shows that Proposition 2.4.1 does not hold for sparse graphs.

**Exercise 2.4.4** (Sparse graphs are not almost regular)☕☕☕ Consider a random graph $G \sim G(n, p)$ with expected degrees $d = o(\log n)$. Show that, with high probability (say, 0.9), $G$ has a vertex with degree $10d$.[2] ☞

[2] We assume here that $10d$ is an integer. There is nothing particular about the factor 10; it could be replaced by any other constant.

Moreover, very sparse graphs, those for which $d = O(1)$, are even farther from regular. The next exercise gives a lower bound on the degrees that matches the upper bound we gave in Exercise 2.4.3.

**Exercise 2.4.5** (Very sparse graphs are far from being regular)☕☕ Consider a random graph $G \sim G(n, p)$ with expected degrees $d = O(1)$. Show that, with high probability (say, 0.9), $G$ has a vertex whose degree is at least of order

$$\frac{\log n}{\log \log n}.$$

## 2.5 Sub-Gaussian Distributions

So far, we have studied concentration inequalities that apply only for Bernoulli random variables $X_i$. It would be useful to extend these results to a wider class of distributions. At the very least we may expect that the normal distribution belongs to this class, since we think of concentration results as quantitative versions of the central limit theorem.

So let us ask which random variables $X_i$ must obey a concentration inequality like Hoeffding's in Theorem 2.2.5, namely

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

If the sum $\sum_{i=1}^{N} a_i X_i$ consists of a single term $a_i X_i$, this inequality reads as

$$\mathbb{P}\left\{|X_i| > t\right\} \leq 2e^{-ct^2}.$$

This gives us an automatic restriction: if we want Hoeffding's inequality to hold, we must assume that the random variables $X_i$ have sub-gaussian tails.

This class of distributions, which we call *sub-gaussian*, deserves special attention. It is sufficiently wide as it contains Gaussian, Bernoulli, and all bounded distributions. And, as we will see shortly, concentration results like Hoeffding's inequality can indeed be proved for all sub-gaussian distributions. This makes the family of sub-gaussian distributions a natural, and in many cases canonical, class where one can develop various results in high-dimensional probability theory and its applications.

We now explore several equivalent approaches to sub-gaussian distributions, examining the behavior of their tails, moments, and moment generating functions. To pave our way, let us recall how these quantities behave for the standard normal distribution.

Let $X \sim N(0, 1)$. Then using (2.3) and symmetry, we obtain the following tail bound:

$$\mathbb{P}\left\{|X| \geq t\right\} \leq 2e^{-t^2/2} \quad \text{for all } t \geq 0. \tag{2.10}$$

(Deduce this formally!) In the next exercise we obtain a bound on the absolute moments and $L^p$ norms of the normal distribution.

**Exercise 2.5.1** (Moments of the normal distribution)☕☕ Show that, for each $p \geq 1$, the random variable $X \sim N(0, 1)$ satisfies

$$\|X\|_{L^p} = (\mathbb{E}\,|X|^p)^{1/p} = \sqrt{2}\Big(\frac{\Gamma((1+p)/2)}{\Gamma(1/2)}\Big)^{1/p}.$$

Deduce that

$$\|X\|_{L^p} = O(\sqrt{p}) \quad \text{as } p \to \infty. \tag{2.11}$$

Finally, a classical formula gives the moment generating function of $X \sim N(0,1)$:

$$\mathbb{E}\exp(\lambda X) = e^{\lambda^2/2} \quad \text{for all } \lambda \in \mathbb{R}. \tag{2.12}$$

### 2.5.1 Sub-Gaussian Properties

Now let $X$ be a general random variable. The following proposition states that the properties we just considered are equivalent – a sub-gaussian tail decay as in (2.10), the growth of moments as in (2.11), and the growth of the moment generating function as in (2.12). The proof of this result is quite useful; it shows how to transform one type of information about random variables into another.

**Proposition 2.5.2** (Sub-gaussian properties) *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.*[3]

(i) The tails of $X$ satisfy

$$\mathbb{P}\{|X| \geq t\} \leq 2\exp(-t^2/K_1^2) \quad \text{for all } t \geq 0.$$

(ii) The moments of $X$ satisfy

$$\|X\|_{L^p} = (\mathbb{E}|X|^p)^{1/p} \leq K_2\sqrt{p} \quad \text{for all } p \geq 1.$$

(iii) The MGF of $X^2$ satisfies

$$\mathbb{E}\exp(\lambda^2 X^2) \leq \exp(K_3^2\lambda^2) \quad \text{for all } \lambda \text{ such that } |\lambda| \leq \frac{1}{K_3}.$$

(iv) The MGF of $X^2$ is bounded at some point, namely

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

*Moreover, if $\mathbb{E}X = 0$ then properties* (i)–(iv) *are also equivalent to the following property.*

(v) The MGF of $X$ satisfies

$$\mathbb{E}\exp(\lambda X) \leq \exp(K_5^2\lambda^2) \quad \text{for all } \lambda \in \mathbb{R}.$$

*Proof* (i) ⇒ (ii) Assume that property (i) holds. By homogeneity and rescaling $X$ to $X/K_1$ we can assume that $K_1 = 1$. Applying the integral identity (Lemma 1.2.1) for $|X|^p$, we obtain

$$\begin{aligned} \mathbb{E}|X|^p &= \int_0^\infty \mathbb{P}\{|X|^p \geq u\}\,du \\ &= \int_0^\infty \mathbb{P}\{|X| \geq t\}\,pt^{p-1}\,dt \quad \text{(by change of variables } u = t^p) \end{aligned}$$

[3] The precise meaning of this equivalence is the following. There exists an absolute constant $C$ such that property $i$ implies property $j$ with parameter $K_j \leq CK_i$ for any two properties $i, j = 1, \ldots, 5$.

$$\begin{aligned}
&\le \int_0^\infty 2e^{-t^2} p t^{p-1}\, dt \quad \text{(by property (i))}\\
&= p\Gamma(p/2) \quad \text{(set } t^2 = s \text{ and use definition of Gamma function)}\\
&\le p(p/2)^{p/2} \quad \text{(since } \Gamma(x) \le x^x \text{ by Stirling's approximation).}
\end{aligned}$$

Taking the $p$th root yields property (ii) with $K_2 \le 2$.

(ii) ⇒ (iii) Assume that property (ii) holds. As before, by homogeneity we may assume that $K_2 = 1$. Recalling the Taylor series expansion of the exponential function, we obtain

$$\mathbb{E}\exp(\lambda^2 X^2) = \mathbb{E}\left(1 + \sum_{p=1}^{\infty} \frac{(\lambda^2 X^2)^p}{p!}\right) = 1 + \sum_{p=1}^{\infty} \frac{\lambda^{2p}\,\mathbb{E}(X^{2p})}{p!}.$$

Property (ii) guarantees that $\mathbb{E}[X^{2p}] \le (2p)^p$, while Stirling's approximation yields $p! \ge (p/e)^p$. Substituting these two bounds, we get

$$\mathbb{E}\exp(\lambda^2 X^2) \le 1 + \sum_{p=1}^{\infty} \frac{(2\lambda^2 p)^p}{(p/e)^p} = \sum_{p=0}^{\infty} (2e\lambda^2)^p = \frac{1}{1 - 2e\lambda^2},$$

provided that $2e\lambda^2 < 1$, in which case the geometric series above converges. To bound this quantity further, we can use the numeric inequality $1/(1-x) \le e^{2x}$, which is valid for $x \in [0, 1/2]$. It follows that

$$\mathbb{E}\exp(\lambda^2 X^2) \le \exp(4e\lambda^2) \quad \text{for all } \lambda \text{ satisfying } |\lambda| \le \frac{1}{2\sqrt{e}}.$$

This yields property (iii) with $K_3 = 1/2\sqrt{e}$.

(iii) ⇒ (iv) is trivial.

(iv) ⇒ (i) Assume that property (iv) holds. As before, we may assume that $K_4 = 1$. Then

$$\begin{aligned}
\mathbb{P}\{|X| \ge t\} &= \mathbb{P}\{e^{X^2} \ge e^{t^2}\}\\
&\le e^{-t^2}\,\mathbb{E}\, e^{X^2} \quad \text{(by Markov's inequality, Proposition 1.2.4)}\\
&\le 2e^{-t^2} \quad \text{(by property (iv)).}
\end{aligned}$$

This proves property (i) with $K_1 = 1$.

To prove the second part of the proposition, we show that (iii) ⇒ (v) and (v) ⇒ (i).

(iii) ⇒ (v) Assume that property (iii) holds; as before we can assume that $K_3 = 1$. Let us use the numeric inequality $e^x \le x + e^{x^2}$, which is valid for all $x \in \mathbb{R}$. Then

$$\begin{aligned}
\mathbb{E}\, e^{\lambda X} &\le \mathbb{E}\left(\lambda X + e^{\lambda^2 X^2}\right)\\
&= \mathbb{E}\, e^{\lambda^2 X^2} \quad \text{(since } \mathbb{E}\, X = 0 \text{ by assumption)}\\
&\le e^{\lambda^2} \quad \text{if } |\lambda| \le 1,
\end{aligned}$$

where in the last line we used property (iii). Thus we have proved property (v) in the range $|\lambda| \le 1$. Now assume that $|\lambda| \ge 1$. Here we can use the numeric inequality $2\lambda x \le \lambda^2 + x^2$, which is valid for all $\lambda$ and $x$. It follows that

$$\begin{aligned} \mathbb{E}\, e^{\lambda X} &\le e^{\lambda^2/2}\, \mathbb{E}\, e^{X^2/2} \le e^{\lambda^2/2} \exp(1/2) \quad \text{(by property (iii))} \\ &\le e^{\lambda^2} \quad \text{(since } |\lambda| \ge 1\text{)}. \end{aligned}$$

This proves property (v) with $K_5 = 1$.

(v) ⇒ (i) Assume that property (v) holds; we can assume that $K_5 = 1$. We will use some ideas from the proof of Hoeffding's inequality (Theorem 2.2.2). Let $\lambda > 0$ be a parameter to be chosen later. Then

$$\begin{aligned} \mathbb{P}\{X \ge t\} &= \mathbb{P}\{e^{\lambda X} \ge e^{\lambda t}\} \\ &\le e^{-\lambda t}\, \mathbb{E}\, e^{\lambda X} \quad \text{(by Markov's inequality)} \\ &\le e^{-\lambda t} e^{\lambda^2} \quad \text{(by property (v))} \\ &= e^{-\lambda t + \lambda^2}. \end{aligned}$$

Optimizing in $\lambda$ and thus choosing $\lambda = t/2$, we conclude that

$$\mathbb{P}\{X \ge t\} \le e^{-t^2/4}.$$

Repeating this argument for $-X$, we also obtain $\mathbb{P}\{X \le -t\} \le e^{-t^2/4}$. Combining these two bounds we conclude that

$$\mathbb{P}\{|X| \ge t\} \le 2e^{-t^2/4}.$$

Thus property (i) holds with $K_1 = 2$. The proposition is proved. ■

**Remark 2.5.3** The constant 2 that appears in some properties in Proposition 2.5.2 does not have any special meaning; it can be replaced by other absolute constants. (Check!)

**Exercise 2.5.4**☕☕ Show that the condition $\mathbb{E}\, X = 0$ is necessary for property (v) to hold.

**Exercise 2.5.5** (On property (iii) in Proposition 2.5.2)☕☕

(a) Show that if $X \sim N(0,1)$, the function $\lambda \mapsto \mathbb{E} \exp(\lambda^2 X^2)$ of $X^2$ is finite only in some bounded neighborhood of zero.

(b) Suppose that some random variable $X$ satisfies $\mathbb{E} \exp(\lambda^2 X^2) \le \exp(K\lambda^2)$ for all $\lambda \in \mathbb{R}$ and some constant $K$. Show that $X$ is a bounded random variable, i.e. $\|X\|_\infty < \infty$.

### *2.5.2 Definition and Examples of Sub-Gaussian Distributions*

**Definition 2.5.6** (Sub-gaussian random variables) A random variable $X$ that satisfies one of the equivalent properties (i)–(iv) in Proposition 2.5.2 is called a *sub-gaussian random variable*. The *sub-gaussian norm* of $X$, denoted $\|X\|_{\psi_2}$, is defined to be the smallest $K_4$ in property (iv). In other words, we define

$$\|X\|_{\psi_2} = \inf \left\{ t > 0 :\ \mathbb{E} \exp(X^2/t^2) \le 2 \right\}. \tag{2.13}$$

**Exercise 2.5.7**☕☕ Check that $\|\cdot\|_{\psi_2}$ is indeed a norm on the space of sub-gaussian random variables.

Let us restate Proposition 2.5.2 in terms of the sub-gaussian norm. It states that every sub-gaussian random variable $X$ satisfies the following bounds:

$$\mathbb{P}\{|X| \geq t\} \leq 2\exp(-ct^2/\|X\|_{\psi_2}^2) \quad \text{for all } t \geq 0; \tag{2.14}$$

$$\|X\|_{L^p} \leq C\|X\|_{\psi_2}\sqrt{p} \quad \text{for all } p \geq 1; \tag{2.15}$$

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

$$\text{if } \mathbb{E}\, X = 0 \text{ then } \mathbb{E}\exp(\lambda X) \leq \exp(C\lambda^2\|X\|_{\psi_2}^2) \quad \text{for all } \lambda \in \mathbb{R}. \tag{2.16}$$

Here $C, c > 0$ are absolute constants. Moreover, up to absolute constant factors, $\|X\|_{\psi_2}$ is the smallest possible number that makes each of these inequalities valid.

**Example 2.5.8** Here are some classical examples of sub-gaussian distributions.

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

$$\|X\|_{\psi_2} \leq C\sigma.$$

(Why?)

(ii) **(Bernoulli)** Let $X$ be a random variable with symmetric Bernoulli distribution (recall Definition 2.2.1). Since $|X| = 1$, it follows that $X$ is a sub-gaussian random variable with

$$\|X\|_{\psi_2} = \frac{1}{\sqrt{\ln 2}}.$$

(iii) **(Bounded)** More generally, any bounded random variable $X$ is sub-gaussian with

$$\|X\|_{\psi_2} \leq C\|X\|_\infty, \tag{2.17}$$

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

**Exercise 2.5.9**☕ Check that the Poisson, exponential, Pareto, and Cauchy distributions are not sub-gaussian.

**Exercise 2.5.10** (Maximum of sub-gaussian)☕☕☕ Let $X_1, X_2, \ldots,$ be an infinite sequence of sub-gaussian random variables which are not necessarily independent. Show that

$$\mathbb{E}\max_i \frac{|X_i|}{\sqrt{1 + \log i}} \leq CK,$$

where $K = \max_i \|X_i\|_{\psi_2}$. Deduce that for every $N \geq 2$ we have

$$\mathbb{E}\max_{i \leq N} |X_i| \leq CK\sqrt{\log N}.$$

**Exercise 2.5.11** (Lower bound)☕☕ Show that the bound in Exercise 2.5.10 is sharp. Let $X_1, X_2, \ldots, X_N$ be independent $N(0, 1)$ random variables. Prove that

$$\mathbb{E}\max_{i \leq N} X_i \geq c\sqrt{\log N}.$$

## 2.6 General Hoeffding and Khintchine Inequalities

After all the work we did in characterizing sub-gaussian distributions in the previous section, we can now easily extend Hoeffding's inequality (Theorem 2.2.2) to general sub-gaussian distributions. But before we do this, let us deduce an important *rotation invariance* property of sums of independent sub-gaussians.

Recall that a sum of independent normal random variables $X_i$ is normal. Indeed, if the $X_i \sim N(0, \sigma_i^2)$ are independent then

$$\sum_{i=1}^N X_i \sim N\Big(0, \sum_{i=1}^N \sigma_i^2\Big). \tag{2.18}$$

This fact is a form of the *rotation invariance* property of the normal distribution, which we will recall in Section 3.3.2 in more detail.

The rotation invariance property extends to general sub-gaussian distributions, albeit up to an absolute constant.

**Proposition 2.6.1** (Sums of independent sub-gaussians) *Let $X_1, \dots, X_N$ be independent mean-zero sub-gaussian random variables. Then $\sum_{i=1}^N X_i$ is also a sub-gaussian random variable, and*

$$\Big\| \sum_{i=1}^N X_i \Big\|_{\psi_2}^2 \le C \sum_{i=1}^N \|X_i\|_{\psi_2}^2,$$

*where $C$ is an absolute constant.*

*Proof* Let us analyze the moment generating function of the sum. For any $\lambda \in \mathbb{R}$, we have

$$\begin{aligned} \mathbb{E} \exp\Big(\lambda \sum_{i=1}^N X_i\Big) &= \prod_{i=1}^N \mathbb{E} \exp(\lambda X_i) \quad \text{(by independence)} \\ &\le \prod_{i=1}^N \exp(C\lambda^2 \|X_i\|_{\psi_2}^2) \quad \text{(by sub-gaussian property (2.16))} \\ &= \exp(\lambda^2 K^2) \quad \text{where } K^2 := C \sum_{i=1}^N \|X_i\|_{\psi_2}^2. \end{aligned}$$

To complete the proof, we just need to recall that the bound on the MGF that we have just proved characterizes sub-gaussian distributions. Indeed, the equivalence of properties (v) and (iv) in Proposition 2.5.2 and Definition 2.5.6 implies that the sum $\sum_{i=1}^N X_i$ is sub-gaussian and

$$\Big\| \sum_{i=1}^N X_i \Big\|_{\psi_2} \le C_1 K$$

where $C_1$ is an absolute constant. The proposition is proved. ■

The approximate rotation invariance property can be restated as a concentration inequality via (2.14):

**Theorem 2.6.2** (General Hoeffding inequality) *Let $X_1, \dots, X_N$ be independent mean-zero sub-gaussian random variables. Then, for every $t \geq 0$, we have*

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

To compare this general result with the specific case for Bernoulli distributions (Theorem 2.2.2), let us apply this result for $a_i X_i$ instead of $X_i$. We obtain a general form of Theorem 2.2.2 for sub-gaussian random variables.

**Theorem 2.6.3** (General Hoeffding inequality) *Let $X_1, \dots, X_N$ be independent mean-zero sub-gaussian random variables, and let $a = (a_1, \dots, a_N) \in \mathbb{R}^N$. Then, for every $t \geq 0$, we have*

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

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

**Exercise 2.6.4**☕ Deduce Hoeffding's inequality for bounded random variables (Theorem 2.2.6) from Theorem 2.6.3, possibly with some other absolute constant instead of 2 in the exponent.

As an application of the general Hoeffding inequality, we can quickly derive the classical Khintchine inequality for the $L^p$ norms of sums of independent random variables.

**Exercise 2.6.5** (Khintchine's inequality)☕☕ Let $X_1, \dots, X_N$ be independent sub-gaussian random variables with zero means and unit variances, and let $a = (a_1, \dots, a_N) \in \mathbb{R}^N$. Prove that for every $p \in [2, \infty)$ we have

$$\left(\sum_{i=1}^{N} a_i^2\right)^{1/2} \leq \left\|\sum_{i=1}^{N} a_i X_i\right\|_{L^p} \leq CK\sqrt{p}\left(\sum_{i=1}^{N} a_i^2\right)^{1/2}$$

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

**Exercise 2.6.6** (Khintchine's inequality for $p = 1$)☕☕☕ Show that, in the setting of Exercise 2.6.5, we have

$$c(K)\left(\sum_{i=1}^{N} a_i^2\right)^{1/2} \leq \left\|\sum_{i=1}^{N} a_i X_i\right\|_{L^1} \leq \left(\sum_{i=1}^{N} a_i^2\right)^{1/2}.$$

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

**Exercise 2.6.7** (Khintchine's inequality for $p \in (0, 2)$)☕☕ State and prove a version of Khintchine's inequality for $p \in (0, 2)$. ☞

### 2.6.1 Centering

In results like Hoeffding's inequality, and in many other results that we will encounter later, we typically assume that the random variables $X_i$ have zero means. If this is not the case, we can always center a variable $X_i$ by subtracting the mean. Let us check that centering does not harm the sub-gaussian property.

First note the following simple centering inequality for the $L^2$ norm:

$$\|X - \mathbb{E}\,X\|_{L^2} \le \|X\|_{L^2}. \tag{2.19}$$

(Check this!) Now let us prove a similar centering inequality for the sub-gaussian norm.

**Lemma 2.6.8** (Centering) *If $X$ is a sub-gaussian random variable then $X - \mathbb{E}\,X$ is sub-gaussian too and*

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

*where $C$ is an absolute constant.*

*Proof* Recall from Exercise 2.5.7 that $\|\cdot\|_{\psi_2}$ is a norm. Thus we can use the triangle inequality and get

$$\|X - \mathbb{E}\,X\|_{\psi_2} \le \|X\|_{\psi_2} + \|\,\mathbb{E}\,X\|_{\psi_2}. \tag{2.20}$$

We only have to bound the second term. Note that, for any constant random variable $a$, we trivially have $\|a\|_{\psi_2} \lesssim |a|$ (recall 2.17).[4] Using this for $a = \mathbb{E}\,X$, we get

$$\begin{aligned}
\|\,\mathbb{E}\,X\|_{\psi_2} &\lesssim |\,\mathbb{E}\,X| \\
&\le \mathbb{E}\,|X| \quad \text{(by Jensen's inequality)} \\
&= \|X\|_1 \\
&\lesssim \|X\|_{\psi_2} \quad \text{(using (2.15) with } p = 1\text{)}.
\end{aligned}$$

Substituting this into (2.20), we complete the proof. ■

**Exercise 2.6.9**☕☕☕ Show that, unlike (2.19), the centering inequality in Lemma 2.6.8 does not hold with $C = 1$.

## 2.7 Sub-Exponential 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. Here is one example. 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

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

[4] In this proof and later, the notation $a \lesssim b$ means that $a \le Cb$ where $C$ is some absolute constant.

Here we find ourselves in a strange situation. On the one hand, $\|g\|_2^2$ is a sum of independent random variables $g_i^2$, so we should expect some concentration to hold. On the other hand, although the $g_i$ are sub-gaussian random variables, the $g_i^2$ are not. Indeed, recalling the behavior of Gaussian tails (Proposition 2.1.2) we have[5]

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

The tails of $g_i^2$ are like those for the exponential distribution and are strictly heavier than sub-gaussian. This prevents us from using Hoeffding's inequality (Theorem 2.6.2) if we want to study the concentration of $\|g\|_2$.

In this section we focus on the class of distributions that have at least an exponential tail decay, and in Section 2.8 we prove an analog of Hoeffding's inequality for them.

Our analysis here will be quite similar to what we did for sub-gaussian distributions in Section 2.5. The following is a version of Proposition 2.5.2 for sub-exponential distributions.

**Proposition 2.7.1** (Sub-exponential properties) *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.*[6]

(i) The tails of $X$ satisfy

$$\mathbb{P}\{|X| \geq t\} \leq 2\exp(-t/K_1) \quad \text{for all } t \geq 0.$$

(ii) The moments of $X$ satisfy

$$\|X\|_{L^p} = (\mathbb{E}\,|X|^p)^{1/p} \leq K_2 p \quad \text{for all } p \geq 1.$$

(iii) The MGF of $|X|$ satisfies

$$\mathbb{E}\exp(\lambda|X|) \leq \exp(K_3\lambda) \quad \text{for all } \lambda \text{ such that } 0 \leq \lambda \leq 1/K_3.$$

(iv) The MGF of $|X|$ is bounded at some point, namely

$$\mathbb{E}\exp(|X|/K_4) \leq 2.$$

*Moreover, if $\mathbb{E}\,X = 0$ then properties (i)–(iv) are also equivalent to the following property.*

(v) The MGF of $X$ satisfies

$$\mathbb{E}\exp(\lambda X) \leq \exp(K_5^2\lambda^2) \quad \text{for all } \lambda \text{ such that } |\lambda| \leq 1/K_5.$$

*Proof* We will prove the equivalence of properties (ii) and (v) only; the reader can check the other implications in Exercise 2.7.2.

[5] Here we have ignored the pre-factor $1/t$, which does not have much effect on the exponent.

[6] The precise meaning of this equivalence is the following. There exists an absolute constant $C$ such that property $i$ implies property $j$ with parameter $K_j \leq CK_i$ for any two properties $i, j = 1, 2, 3, 4$.

(ii) ⇒ (v) Without loss of generality we may assume that $K_2 = 1$. (Why?) Expanding the exponential function in a Taylor series, we obtain

$$\mathbb{E}\exp(\lambda X) = \mathbb{E}\left(1 + \lambda X + \sum_{p=2}^{\infty} \frac{(\lambda X)^p}{p!}\right) = 1 + \sum_{p=2}^{\infty} \frac{\lambda^p \, \mathbb{E}\, X^p}{p!},$$

where we have used the assumption that $\mathbb{E}\, X = 0$. Property (ii) guarantees that $\mathbb{E}\; X^p \le p^p$, while Stirling's approximation yields $p! \ge (p/e)^p$. Substituting these two bounds, we obtain

$$\mathbb{E}\exp(\lambda X) \le 1 + \sum_{p=2}^{\infty} \frac{(\lambda p)^p}{(p/e)^p} = 1 + \sum_{p=2}^{\infty} (e\lambda)^p = 1 + \frac{(e\lambda)^2}{1 - e\lambda},$$

provided that $|e\lambda| < 1$, in which case the geometric series above converges. Moreover, if $|e\lambda| \le 1/2$ then we can further bound the above quantity by

$$1 + 2e^2\lambda^2 \le \exp(2e^2\lambda^2).$$

Summarizing, we have shown that

$$\mathbb{E}\exp(\lambda X) \le \exp(2e^2\lambda^2) \quad \text{for all } \lambda \text{ satisfying } |\lambda| \le \frac{1}{2e}.$$

This yields property (v) with $K_5 = 1/(2e)$.

(v) ⇒ (ii) Without loss of generality, we can assume that $K_5 = 1$. We will use the numeric inequality

$$|x|^p \le p^{\,p}(e^x + e^{-x}),$$

which is valid for all $x \in \mathbb{R}$ and $p > 0$. (Check it by dividing both sides by $p^p$ and taking $p$th roots.) Substituting $x = X$ and taking expectations, we get

$$\mathbb{E}\,|X|^p \le p^{\,p}\big(\mathbb{E}\, e^X + \mathbb{E}\, e^{-X}\big).$$

Property (v) gives $\mathbb{E}\, e^X \le 1$ and $\mathbb{E}\, e^{-X} \le 1$. Thus

$$\mathbb{E}\,|X|^p \le 2p^{\,p}.$$

This yields property (ii) with $K_2 = 2$. ■

**Exercise 2.7.2**☕☕ Prove the equivalence of properties (i)–(iv) in Proposition 2.7.1 by modifying the proof of Proposition 2.5.2.

**Exercise 2.7.3**☕☕☕ More generally, consider the class of distributions whose tail decay is of the type $\exp(-ct^\alpha)$ or faster. Here $\alpha = 2$ corresponds to sub-gaussian distributions and $\alpha = 1$ to sub-exponential distributions. State and prove a version of Proposition 2.7.1 for such distributions.

**Exercise 2.7.4**☕ Argue that the bound in property (iii) in Proposition 2.7.1 cannot be extended to all $\lambda$ such that $|\lambda| \le 1/K_3$.

**Definition 2.7.5** (Sub-exponential random variables) A random variable $X$ that satisfies one of the equivalent properties (i)–(iv) in Proposition 2.7.1 is called a *sub-exponential random variable*. The *sub-exponential norm* of $X$, denoted $\|X\|_{\psi_1}$, is defined to be the smallest $K_3$ in property (iii). In other words,

$$\|X\|_{\psi_1} = \inf\{t > 0 \colon \ \mathbb{E}\exp(|X|/t) \le 2\}. \tag{2.21}$$

Sub-gaussian and sub-exponential distributions are closely related. First, any sub-gaussian distribution is clearly sub-exponential. (Why?) Second, the square of a sub-gaussian random variable is sub-exponential:

**Lemma 2.7.6** (Sub-exponential is sub-gaussian squared) *A random variable $X$ is sub-gaussian if and only if $X^2$ is sub-exponential. Moreover,*

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

*Proof* This follows easily from the definition. Indeed, $\|X^2\|_{\psi_1}$ is the infimum of the numbers $K > 0$ satisfying $\mathbb{E}\exp(X^2/K) \le 2$, while $\|X\|_{\psi_2}$ is the infimum of the numbers $L > 0$ satisfying $\mathbb{E}\exp(X^2/L^2) \le 2$. So these two become the same definition with $K = L^2$. ■

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

**Lemma 2.7.7** (The product of sub-gaussians is sub-exponential) *Let $X$ and $Y$ be sub-gaussian random variables. Then $XY$ is sub-exponential. Moreover,*

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

*Proof* Without loss of generality we may assume that $\|X\|_{\psi_2} = \|Y\|_{\psi_2} = 1$. (Why?) The lemma claims that if

$$\mathbb{E}\exp(X^2) \le 2 \quad \text{and} \quad \mathbb{E}\exp(Y^2) \le 2 \tag{2.22}$$

then $\mathbb{E}\exp(|XY|) \le 2$. To prove this, let us use the elementary form of Young's inequality, which states that

$$ab \le \frac{a^2}{2} + \frac{b^2}{2} \quad \text{for } a, b \in \mathbb{R}.$$

It yields

$$\begin{aligned}
\mathbb{E}\exp(|XY|) &\le \mathbb{E}\exp\Big(\frac{X^2}{2} + \frac{Y^2}{2}\Big) \quad \text{(by Young's inequality)}\\
&= \mathbb{E}\Big(\exp\Big(\frac{X^2}{2}\Big)\exp\Big(\frac{Y^2}{2}\Big)\Big)\\
&\le \frac{1}{2}\,\mathbb{E}\big(\exp(X^2) + \exp(Y^2)\big) \quad \text{(by Young's inequality)}\\
&= \frac{1}{2}(2+2) = 2 \quad \text{(by assumption (2.22))}.
\end{aligned}$$

The proof is complete. ■

**Example 2.7.8** Let us mention a few examples of sub-exponential random variables. As we have just learned, all sub-gaussian random variables and their squares are sub-exponential; for example, $g^2$ for $g \sim N(\mu, \sigma)$. Apart from that, sub-exponential distributions include the exponential and Poisson distributions. Recall that $X$ has an *exponential distribution* with rate $\lambda > 0$, denoted $X \sim \operatorname{Exp}(\lambda)$, if $X$ is a non-negative random variable with tails

$$\mathbb{P}\left\{X \geq t\right\} = e^{-\lambda t} \quad \text{for } t \geq 0.$$

The mean, standard deviation, and sub-exponential norm of $X$ are all of order $1/\lambda$:

$$\mathbb{E}\, X = \frac{1}{\lambda}, \quad \operatorname{Var}(X) = \frac{1}{\lambda^2}, \quad \|X\|_{\psi_1} = \frac{C}{\lambda}.$$

(Check this!)

**Remark 2.7.9** (MGF near the origin) You may be surprised to see the same bound on the MGF near the origin for sub-gaussian and sub-exponential distributions. (Compare property (v) in Propositions 2.5.2 and 2.7.1.) This should not be very surprising, though: this kind of local bound is expected from a *general* random variable $X$ with mean zero and unit variance. To see this, assume for simplicity that $X$ is bounded. The MGF of $X$ can be approximated using the first two terms of a Taylor expansion:

$$\mathbb{E} \exp(\lambda X) \approx \mathbb{E}\left(1 + \lambda X + \frac{\lambda^2 X^2}{2} + o(\lambda^2 X^2)\right) = 1 + \frac{\lambda^2}{2} \approx e^{\lambda^2/2}$$

as $\lambda \to 0$. For the standard *normal* distribution $N(0,1)$, this approximation becomes an equality; see (2.12). For *sub-gaussian* distributions, Proposition 2.5.2 says that a bound like this holds for all $\lambda$ and that this characterizes sub-gaussian distributions. And, for *sub-exponential* distributions, Proposition 2.7.1 says that this bound holds for small $\lambda$ and that this characterizes sub-exponential distributions. For larger $\lambda$, no general bound can exist for sub-exponential distributions: indeed, for the *exponential* random variable $X \sim \operatorname{Exp}(1)$, the MGF is infinite for $\lambda \geq 1$. (Check this!)

**Exercise 2.7.10** (Centering)☕ Prove an analog of the centering lemma 2.6.8 for sub-exponential random variables $X$:

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

### *2.7.1 A More General View: Orlicz Spaces*

Sub-gaussian distributions can be introduced within the more general framework of *Orlicz spaces*. A function $\psi\colon [0,\infty) \to [0,\infty)$ is called an *Orlicz function* if $\psi$ is convex, increasing, and satisfies

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

For a given Orlicz function $\psi$, the Orlicz norm of a random variable $X$ is defined as

$$\|X\|_{\psi} := \inf\left\{t > 0\colon\ \mathbb{E}\, \psi(|X|/t) \leq 1\right\}.$$

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

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

**Exercise 2.7.11**☕☕ Show that $\|X\|_\psi$ is indeed a norm on the space $L_\psi$.

It can also be shown that $L_\psi$ is complete and thus a Banach space.

**Example 2.7.12** ($L^p$ space) Consider the function

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

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

**Example 2.7.13** ($L_{\psi_2}$ 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 in (2.13). The corresponding Orlicz space $L_{\psi_2}$ consists of all sub-gaussian random variables.

**Remark 2.7.14** We can easily locate $L_{\psi_2}$ in the hierarchy of classical $L^p$ spaces:

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

The first inclusion follows from property (ii) of Proposition 2.5.2, and the second inclusion from the bound (2.17). Thus the space of sub-gaussian random variables $L_{\psi_2}$ is smaller than all the $L^p$ spaces, but it is still larger than the space of bounded random variables $L^\infty$.

## 2.8 Bernstein's Inequality

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

**Theorem 2.8.1** (Bernstein's inequality) *Let $X_1, \ldots, X_N$ be independent mean-zero sub-exponential random variables. Then, for every $t \geq 0$, we have*

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

*where $c > 0$ is an absolute constant.*

*Proof* We begin the proof in the same way as we argued about other concentration inequalities for $S = \sum_{i=1}^N X_i$, e.g. Theorems 2.2.2 and 2.3.1. Multiply both sides of the inequality $S \geq t$ by a parameter $\lambda$, exponentiate, and then use Markov's inequality and independence. This leads to the bound (2.7), which is

$$\mathbb{P}\left\{S \ge t\right\} \le e^{-\lambda t} \prod_{i=1}^{N} \mathbb{E}\exp(\lambda X_i). \tag{2.23}$$

To bound the MGF of each term $X_i$, we use property (v) in Proposition 2.7.1. It says that if $\lambda$ is small enough that

$$|\lambda| \le \frac{c}{\max_i \|X_i\|_{\psi_1}} \tag{2.24}$$

then $\mathbb{E}\exp(\lambda X_i) \le \exp\left(C\lambda^2 \|X_i\|_{\psi_1}^2\right)$.[7] Substituting this into (2.23), we obtain

$$\mathbb{P}\{S \ge t\} \le \exp\left(-\lambda t + C\lambda^2\sigma^2\right), \quad \text{where } \sigma^2 = \sum_{i=1}^{N} \|X_i\|_{\psi_1}^2.$$

Now we minimize this expression in $\lambda$ subject to the constraint (2.24). The optimal choice is

$$\lambda = \min\Big(\frac{t}{2C\sigma^2}, \frac{c}{\max_i \|X_i\|_{\psi_1}}\Big),$$

for which we obtain

$$\mathbb{P}\{S \ge t\} \le \exp\Big(-\min\Big(\frac{t^2}{4C\sigma^2}, \frac{ct}{2\max_i \|X_i\|_{\psi_1}}\Big)\Big).$$

Repeating this argument for $-X_i$ instead of $X_i$, we obtain the same bound for $\mathbb{P}\{-S \ge t\}$. A combination of these two bounds completes the proof. ■

To put Theorem 2.8.1 in a more convenient form, let us apply it for $a_i X_i$ instead of $X_i$.

**Theorem 2.8.2** (Bernstein's inequality) *Let $X_1, \dots, X_N$ be independent, mean-zero sub-exponential random variables, and let $a = (a_1, \dots, a_N) \in \mathbb{R}^N$. Then, for every $t \ge 0$, we have*

$$\mathbb{P}\left\{\left|\sum_{i=1}^{N} a_i X_i\right| \ge t\right\} \le 2\exp\Big(-c\min\Big(\frac{t^2}{K^2\|a\|_2^2}, \frac{t}{K\|a\|_\infty}\Big)\Big)$$

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

In the special case where $a_i = 1/N$, we obtain a form of Bernstein's inequality for averages:

**Corollary 2.8.3** (Bernstein's inequality) *Let $X_1, \dots, X_N$ be independent mean-zero, sub-exponential random variables. Then, for every $t \ge 0$, we have*

$$\mathbb{P}\left\{\left|\frac{1}{N}\sum_{i=1}^{N} X_i\right| \ge t\right\} \le 2\exp\Big(-c\min\Big(\frac{t^2}{K^2}, \frac{t}{K}\Big)N\Big)$$

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

[7] Recall that by Proposition 2.7.1 and the definition of the sub-exponential norm, property (v) holds for a value of $K_5$ that is within an absolute constant factor of $\|X\|_{\psi_1}$.

This result can be considered as a quantitative form of the *law of large numbers* for the averages $\frac{1}{N}\sum_{i=1}^N X_i$.

Let us compare Bernstein's inequality (Theorem 2.8.1) with Hoeffding's inequality (Theorem 2.6.2). The obvious difference is that Bernstein's bound has *two tails*, as it would if the sum $S_N = \sum X_i$ were a mixture of sub-gaussian and sub-exponential distributions. The sub-gaussian tail is of course expected from the central limit theorem. But the sub-exponential tails of the terms $X_i$ are too heavy to be able to produce a sub-gaussian tail everywhere, so a sub-exponential tail should be expected, too. In fact, the sub-exponential tail in Theorem 2.8.1 is produced by a *single term* $X_i$ in the sum, the one with the maximal sub-exponential norm. Indeed, this term alone has a tail of magnitude $\exp(-ct/\|X_i\|_{\psi_1})$.

We have already seen a similar mixture of two tails, one for small deviations and the other for large deviations, in our analysis of Chernoff's inequality; see Remark 2.3.7. To put Bernstein's inequality in the same perspective, let us normalize the sum as in the central limit theorem and apply Theorem 2.8.2. We obtain[8]

$$\mathbb{P}\left\{\left|\frac{1}{\sqrt{N}}\sum_{i=1}^N X_i\right| \ge t\right\} \le \begin{cases} 2\exp(-ct^2), & t \le C\sqrt{N}, \\ 2\exp(-t\sqrt{N}), & t \ge C\sqrt{N}. \end{cases}$$

Thus, in the *small-deviation* regime, where $t \le C\sqrt{N}$, we have the same 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. For *large deviations*, where $t \ge C\sqrt{N}$, the sum has a heavier, *sub-exponential*, tail bound, which can be due to the contribution of a single term $X_i$. We illustrate this in Figure 2.3.

Let us mention the following strengthening of Bernstein's inequality under the stronger assumption that the random variables $X_i$ are bounded.

**Theorem 2.8.4** (Bernstein's inequality for bounded distributions) *Let $X_1, \dots, X_N$ be independent mean-zero random variables such that $|X_i| \le K$ all $i$. Then, for every $t \ge 0$, we have*

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

*Here $\sigma^2 = \sum_{i=1}^N \mathbb{E} X_i^2$ is the variance of the sum.*

We leave the proof of this theorem to the next two exercises.

**Exercise 2.8.5** (A bound on MGF)☕☕ Let $X$ be a random variable such that $|X| \le K$. Prove the following bound on the MGF of $X$:

$$\mathbb{E}\exp(\lambda X) \le \exp(g(\lambda)\,\mathbb{E} X^2) \quad \text{where } g(\lambda) = \frac{\lambda^2/2}{1 - |\lambda|K/3},$$

provided that $|\lambda| < 3/K$. ☞

[8] For simplicity, we have suppressed the dependence on $K$ here by allowing the constants $c, C$ to depend on $K$.

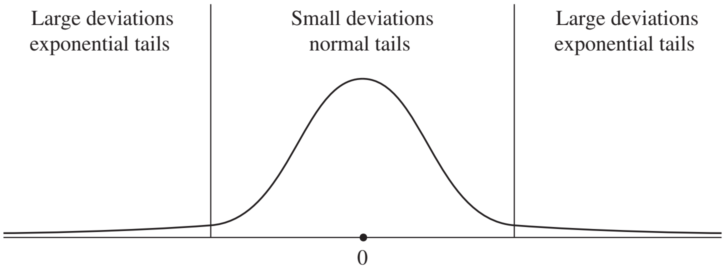

**Figure 2.3** Bernstein's inequality for a sum of sub-exponential random variables gives a mixture of two tails: sub-gaussian for small deviations and sub-exponential for large deviations.

**Exercise 2.8.6**☕☕ Deduce Theorem 2.8.4 from the bound in Exercise 2.8.5. ☞

## 2.9 Notes

The topic of concentration inequalities is very wide, and we will continue to examine it in Chapter 5. We refer the reader to [8, Appendix A], [148, Chapter 4], [126], [29], [76, Chapter 7], [11, Section 3.5.4], [170, Chapter 1], and [14, Chapter 4] for various versions of Hoeffding's, Chernoff's, and Bernstein's inequalities and related results.

Proposition 2.1.2 on the tails of the normal distribution is taken from [70, Theorem 1.4]. The proof of the Berry–Esseen central limit theorem (Theorem 2.1.3) with an extra factor 3 on the right-hand side can be found e.g. in [70, Section 2.4.d]; the best currently known factor is $\approx 0.47$ [116].

It is worthwhile mentioning two important concentration inequalities that were omitted in this chapter. One is the *bounded differences inequality*, also called *McDiarmid's inequality*, which works not only for sums but for general functions of independent random variables. It is a generalization of Hoeffding's inequality (Theorem 2.2.6).

**Theorem 2.9.1** (Bounded differences inequality) *Let $X_1, \dots, X_N$ be independent random variables.*[9] *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a measurable function. Assume that the value of $f(x)$ can change by at most $c_i > 0$ under an arbitrary change*[10] *of a single coordinate of $x \in \mathbb{R}^n$. Then, for any $t > 0$, we have*

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

*where $X = (X_1, \dots, X_n)$.*

Another result worth mentioning is *Bennett's inequality*, which can be regarded as a generalization of Chernoff's inequality.

[9] The theorem remains valid if the random variables $X_i$ take values in an abstract set $\mathcal{X}$ and $f : \mathcal{X} \to \mathbb{R}$.

[10] This means that for any index $i$ and any $x_1, \dots, x_n, x_i'$, we have
$|f(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n) - f(x_1, \dots, x_{i-1}, x_i', x_{i+1}, \dots, x_n)| \leq c_i$.

**Theorem 2.9.2** (Bennett's inequality) *Let $X_1, \dots, X_N$ be independent random variables. Assume that $|X_i - \mathbb{E} X_i| \le K$ almost surely for every $i$. Then, for any $t > 0$, we have*

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

*where $\sigma^2 = \sum_{i=1}^{N} \operatorname{Var}(X_i)$ is the variance of the sum, and $h(u) = (1+u)\log(1+u) - u$.*

In the small-deviation regime, where $u := Kt/\sigma^2 \ll 1$, we have asymptotically $h(u) \approx u^2$ and Bennett's inequality gives approximately the Gaussian tail bound $\approx \exp(-t^2/\sigma^2)$. In the large-deviation regime, say where $u \gg Kt/\sigma^2 \ge 2$, we have $h(u) \ge \frac{1}{2} u \log u$, and Bennett's inequality gives a Poisson-like tail $(\sigma^2/Kt)^{t/2K}$.

Both the bounded differences inequality and Bennett's inequality can be proved by the same general method as Hoeffding's inequality (Theorem 2.2.2) and Chernoff's inequality (Theorem 2.3.1), namely by bounding the moment generating function of the sum. This method was pioneered by Sergei Bernstein in the 1920s and 1930s. Our presentation of Chernoff's inequality in Section 2.3 mostly follows [148, Chapter 4].

Section 2.4 scratches the surface of the rich theory of *random graphs*. The books [25, 105] offer a comprehensive introduction to the random graph theory.

The presentation in Sections 2.5–2.8 mostly follows [216]; see [76, Chapter 7] for some more elaborate results. For sharp versions of Khintchine's inequalities in Exercises 2.6.5–2.6.7 and related results, see e.g. [189, 93, 114, 151].