CHAPTER 2

Basic tail and concentration bounds

In a variety of settings, it is of interest to obtain bounds on the tails of a random variable, or two-sided inequalities that guarantee that a random variable is close to its mean or median. In this chapter, we explore a number of elementary techniques for obtaining both deviation and concentration inequalities. It is an entrypoint to more advanced literature on large deviation bounds and concentration of measure.

■ 2.1 Classical bounds

One way in which to control a tail probability $\mathbb{P}[X \geq t]$ is by controlling the moments of the random variable X. Gaining control of higher-order moments leads to correspondingly sharper bounds on tail probabilities, ranging from Markov's inequality (which requires only existence of the first moment) to the Chernoff bound (which requires existence of the moment generating function).

■ 2.1.1 From Markov to Chernoff

The most elementary tail bound is *Markov's inequality*: given a non-negative random variable X with finite mean, we have

$$\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}[X]}{t} \quad \text{for all } t > 0. \tag{2.1}$$

For a random variable X that also has a finite variance, we have *Chebyshev's inequality*:

$$\mathbb{P}\big[|X - \mu| \geq t\big] \leq \frac{\operatorname{var}(X)}{t^2} \quad \text{for all } t > 0. \tag{2.2}$$

Note that this is a simple form of concentration inequality, guaranteeing that X is close to its mean μ whenever its variance is small. Chebyshev's inequality follows by applying Markov's inequality to the non-negative random variable $Y = (X - \mathbb{E}[X])^2$. Both Markov's and Chebyshev's inequality are sharp, meaning that they cannot be improved in general (see Exercise 2.1).

There are various extensions of Markov's inequality applicable to random variables with higher-order moments. For instance, whenever X has a central moment of order k, an application of Markov's inequality to the random variable $|X-\mu|^k$ yields that

$$\mathbb{P}[|X-\mu| \geq t] \leq \frac{\mathbb{E}[|X-\mu|^k]}{t^k} \quad \text{for all } t > 0. \tag{2.3}$$

Of course, the same procedure can be applied to functions other than polynomials $|X-\mu|^k$. For instance, suppose that the random variable X has a moment generating function in a neighborhood of zero, meaning that that there is some constant $b > 0$ such that the function $\varphi(\lambda) = \mathbb{E}[e^{\lambda(X-\mu)}]$ exists for all $\lambda \leq |b|$. In this case, for any $\lambda \in [0, b]$, we may apply Markov's inequality to the random variable $Y = e^{\lambda(X-\mu)}$, thereby obtaining the upper bound

$$\mathbb{P}[(X-\mu) \geq t] = \mathbb{P}[e^{\lambda(X-\mu)} \geq e^{\lambda t}] \;\leq \frac{\mathbb{E}[e^{\lambda(X-\mu)}]}{e^{\lambda t}}. \tag{2.4}$$

Optimizing our choice of λ so as to obtain the tightest result yields the *Chernoff bound*—namely, the inequality

$$\log \mathbb{P}[(X-\mu) \geq t] \leq - \sup_{\lambda \in [0,b]} \big\{\lambda t - \log \mathbb{E}[e^{\lambda(X-\mu)}]\big\}. \tag{2.5}$$

As we explore in Exercise 2.3, the moment bound (2.3) with the optimal choice of k is never worse than the bound (2.5) based on the moment-generating function. Nonetheless, the Chernoff bound is most widely used in practice, possibly due to the ease of manipulating moment generating functions. Indeed, a variety of important tail bounds can be obtained as particular cases of inequality (2.5), as we discuss in examples to follow.

■ 2.1.2 Sub-Gaussian variables and Hoeffding bounds

The form of tail bound obtained via the the Chernoff approach depends on the growth rate of the moment generating function. Accordingly, in the study of tail bounds, it is natural to classify random variables in terms of their moment generating functions. For reasons to become clear in the sequel, the simplest type of behavior is known as sub-Gaussian. In order to motivate this notion, let us illustrate the use of the Chernoff bound (2.5) in deriving tail bounds for a Gaussian variable.

Example 2.1 (Gaussian tail bounds). Let $X \sim \mathcal{N}(\mu, \sigma^2)$ be a Gaussian random variable with mean μ and variance σ^2. By a straightforward calculation, we find that X has the moment generating function

$$\mathbb{E}[e^{\lambda X}] = e^{\mu\lambda + \sigma^2\lambda^2/2}, \qquad \text{valid for all } \lambda \in \mathbb{R}. \tag{2.6}$$

Substituting this expression into the optimization problem defining the optimized Chernoff bound (2.5), we obtain

$$\sup_{\lambda \in \mathbb{R}} \{\lambda t - \log \mathbb{E}[e^{\lambda(X-\mu)}]\} = \sup_{\lambda \in \mathbb{R}} \{\lambda t - \frac{\lambda^2 \sigma^2}{2}\} \; = \; \frac{t^2}{2\sigma^2},$$

where we have taken derivatives in order to find the optimum of this quadratic function. Returning to the Chernoff bound (2.5), we conclude that any $\mathcal{N}(\mu, \sigma^2)$ random variable satisfies the *upper deviation inequality*

$$\mathbb{P}[X \geq \mu + t] \leq e^{-\frac{t^2}{2\sigma^2}} \quad \text{for all } t \geq 0. \tag{2.7}$$

In fact, this bound is sharp up to polynomial-factor corrections, as shown by the result of Exercise 2.2. ♣

Motivated by the structure of this example, we are led to introduce the following definition.

Definition 2.1. A random variable X with mean $\mu = \mathbb{E}[X]$ is *sub-Gaussian* if there is a positive number σ such that

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\sigma^2 \lambda^2/2} \qquad \text{for all } \lambda \in \mathbb{R}. \tag{2.8}$$

The constant σ is referred to as the *sub-Gaussian parameter*; for instance, we say that X is sub-Gaussian with parameter σ when the condition (2.8) holds. Naturally, any Gaussian variable with variance σ^2 is sub-Gaussian with parameter σ, as should be clear from the calculation described in Example 2.1. In addition, as we will see in the examples and exercises to follow, a large number of non-Gaussian random variables also satisfy the condition (2.8).

The condition (2.8), when combined with the Chernoff bound as in Example 2.1, shows that if X is sub-Gaussian with parameter σ, then it satisfies the *upper deviation inequality* (2.7). Moreover, by the symmetry of the definition, the variable $-X$ is sub-Gaussian if and only if X is sub-Gaussian, so that we also have the *lower deviation inequality* $\mathbb{P}[X \leq \mu - t] \leq e^{-\frac{t^2}{2\sigma^2}}$, valid for all $t \geq 0$. Combining the pieces, we conclude that any sub-Gaussian variable satisfies the *concentration inequality*

$$\mathbb{P}[|X - \mu| \geq t] \leq 2\, e^{-\frac{t^2}{2\sigma^2}} \quad \text{for all } t \in \mathbb{R}. \tag{2.9}$$

Let us consider some examples of sub-Gaussian variables that are non-Gaussian.

Example 2.2 (Rademacher variables)**.** A Rademacher random variable ε takes the values $\{-1, +1\}$ equiprobably. We claim that it is sub-Gaussian with parameter $\sigma = 1$. By taking expectations and using the power series expansion for the exponential, we obtain

$$\begin{aligned}
\mathbb{E}[e^{\lambda \varepsilon}] &= \frac{1}{2}\Big\{e^{-\lambda} + e^{\lambda}\Big\} \\
&= \frac{1}{2}\Big\{\sum_{k=0}^{\infty} \frac{(-\lambda)^k}{k!} + \sum_{k=0}^{\infty} \frac{(\lambda)^k}{k!}\Big\} \\
&= \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{(2k)!} \\
&\leq 1 + \sum_{k=1}^{\infty} \frac{\lambda^{2k}}{2^k \, k!} \\
&= e^{\lambda^2/2},
\end{aligned}$$

which shows that ε is sub-Gaussian with parameter $\sigma = 1$ as claimed. ♣

We now generalize the preceding example to show that any bounded random variable is also sub-Gaussian.

Example 2.3 (Bounded random variables)**.** Let X be zero-mean, and supported on some interval $[a, b]$. Letting X' be an independent copy, for any $\lambda \in \mathbb{R}$, we have

$$\begin{aligned}
\mathbb{E}_X[e^{\lambda X}] &= \mathbb{E}_X\big[e^{\lambda (X - \mathbb{E}_{X'}[X'])}\big] \\
&\leq \mathbb{E}_{X,X'}\big[e^{\lambda(X - X')}\big],
\end{aligned}$$

where the inequality follows the convexity of the exponential, and Jensen's inequality. Letting ε be an independent Rademacher variable, note that the distribution of $(X - X')$ is the same as that of $\varepsilon(X - X')$, so that we have

$$\mathbb{E}_{X,X'}\big[e^{\lambda(X-X')}\big] = \mathbb{E}_{X,X'}\Big[\mathbb{E}_{\varepsilon}\big[e^{\lambda \varepsilon (X-X')}\big]\Big] \stackrel{(i)}{\leq} \mathbb{E}_{X,X'}\big[e^{\frac{\lambda^2 (X-X')^2}{2}}\big],$$

where step (i) follows from the result of Example 2.2, applied conditionally with (X, X') held fixed. Since $|X - X'| \leq b - a$, we are guaranteed that

$$\mathbb{E}_{X,X'}\big[e^{\frac{\lambda^2 (X-X')^2}{2}}\big] \leq e^{\frac{\lambda^2 \, (b-a)^2}{2}}.$$

Putting together the pieces, we have shown that X is sub-Gaussian with parameter at most $\sigma = b - a$. This result is useful but can be sharpened. In Exercise 2.4, we work through a more involved argument to show that X is sub-Gaussian with parameter at most $\sigma = \frac{b-a}{2}$.

Remark: The technique used in Example 2.3 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, as will be seen in later chapters.

Just as the property of Gaussianity is preserved by linear operations so is the property of sub-Gaussianity. For instance, if X_1, X_2 are independent sub-Gaussian variables with parameters σ_1^2 and σ_2^2, then $X_1 + X_2$ is sub-Gaussian with parameter $\sigma_1^2 + \sigma_2^2$. See Exercise 2.13 for verification of this fact, and related properties. As a consequence of this fact and the basic sub-Gaussian tail bound (2.7), we obtain an important result, applicable to sums of independent sub-Gaussian random variables, and known as the *Hoeffding bound*:

Proposition 2.1 (Hoeffding bound)**.** Suppose that the variables X_i, $i = 1, \ldots, n$ are independent, and X_i has mean μ_i and sub-Gaussian parameter σ_i. Then for all $t \geq 0$, we have

$$\mathbb{P}\Big[\sum_{i=1}^n (X_i - \mu_i) \geq t\Big] \leq \exp\Big\{-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\Big\}. \tag{2.10}$$

The Hoeffding bound is often stated only for the special case of bounded random variables. In particular, if $X_i \in [a, b]$ for all $i = 1, 2, \ldots, n$, then from the result of Exercise 2.4, it is sub-Gaussian with parameter $\sigma = \frac{b-a}{2}$, so that we obtain the bound

$$\mathbb{P}\Big[\sum_{i=1}^n (X_i - \mu_i) \geq t\Big] \leq e^{-\frac{2t^2}{n\,(b-a)^2}}.$$

Although the Hoeffding bound is often stated in this form, the basic idea applies somewhat more generally to sub-Gaussian variables, as we have given here.

We conclude our discussion of sub-Gaussianity with a result that provides three different characterizations of sub-Gaussian variables. First, the most direct way in which to establish sub-Gaussianity is by computing or bounding the moment generating function, as we have done in Example 2.1. A second intuition is that any sub-Gaussian variable is dominated in a certain sense by a Gaussian variable. Third, sub-Gaussianity also follows by having suitably tight control on the moments of the random variable. The following result shows that all three notions are equivalent in a precise sense.

Theorem 2.1 (Equivalent characterizations of sub-Gaussian variables)**.** Given any zero-mean random variable X, the following properties are equivalent:

(I) There is a constant σ such that $\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\lambda^2 \sigma^2}{2}}$ for all $\lambda \in \mathbb{R}$.

(II) There is a constant $c \geq 1$ and Gaussian random variable $Z \sim \mathcal{N}(0, \tau^2)$ such that

$$\mathbb{P}[|X| \geq s] \leq c\, \mathbb{P}[|Z| \geq s] \quad \text{for all } s \geq 0. \tag{2.11}$$

(III) There exists a number $\theta \geq 0$ such that

$$\mathbb{E}[X^{2k}] \leq \frac{(2k)!}{2^k k!}\, \theta^{2k} \quad \text{for all } k = 1, 2, \ldots. \tag{2.12}$$

(IV) We have

$$\mathbb{E}[e^{\frac{\lambda X^2}{2\sigma^2}}] \leq \frac{1}{\sqrt{1-\lambda}} \quad \text{for all } \lambda \in [0, 1). \tag{2.13}$$

See Appendix A for the proof of these equivalences.

■ 2.1.3 Sub-exponential variables and Bernstein bounds

The notion of sub-Gaussianity is fairly restrictive, so that it is natural to consider various relaxations of it. Accordingly, we now turn to the class of sub-exponential variables, which are defined by a slightly milder condition on the moment generating function:

Definition 2.2. A random variable X with mean $\mu = \mathbb{E}[X]$ is *sub-exponential* if there are non-negative parameters (ν, b) such that

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{\nu^2 \lambda^2}{2}} \qquad \text{for all } |\lambda| < \tfrac{1}{b}. \tag{2.14}$$

It follows immediately from this definition that that any sub-Gaussian variable is also sub-exponential—in particular, with $\nu = \sigma$ and $b = 0$, where we interpret $1/0$ as being the same as $+\infty$. However, the converse statement is not true, as shown by the following calculation:

Example 2.4 (Sub-exponential but not sub-Gaussian)**.** Let $Z \sim \mathcal{N}(0, 1)$, and consider the random variable $X = Z^2$. For $\lambda < \frac{1}{2}$, we have

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

For $\lambda > 1/2$, the moment generating function does not exist, showing that X is not sub-Gaussian.

As will be seen momentarily, the existence of the moment generating function in a neighborhood of zero is actually an equivalent definition of a sub-exponential variable. Let us verify directly that condition (2.14) is satisfied. Following some calculus, it can be verified that

$$\frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \leq e^{2\lambda^2} \;=\; e^{4\lambda^2/2}, \qquad \text{for all } |\lambda| < 1/4, \tag{2.15}$$

which shows that X is sub-exponential with parameters $(\nu, b) = (2, 4)$. ♣

As with sub-Gaussianity, the control (2.14) on the moment generating function, when combined with the Chernoff technique, yields deviation and concentration inequalities for sub-exponential variables. When t is small enough, these bounds are sub-Gaussian in nature (i.e. with the exponent quadratic in t), whereas for larger t, the exponential component of the bound scales linearly in t. We summarize in the following:

Proposition 2.2 (Sub-exponential tail bound)**.** Suppose that X is sub-exponential with parameters (ν, b). Then

$$\mathbb{P}[X \geq \mu + t] \leq \begin{cases} e^{-\frac{t^2}{2\nu^2}} & \text{if } 0 \leq t \leq \frac{\nu^2}{b}, \text{ and} \\ e^{-\frac{t}{2b}} & \text{for } t > \frac{\nu^2}{b}. \end{cases}$$

As with the Hoeffding inequality, similar bounds apply to the left-sided event $X \leq \mu - t$, as well as the two-sided event $|X - \mu| \geq t$, with an additional factor of two in the latter case.

Proof. By re-centering as needed, we may assume without loss of generality that $\mu = 0$. We follow the usual Chernoff-type approach: combining it with the definition (2.14) of a sub-exponential variable yields the upper bound

$$\mathbb{P}[X \geq t] \leq e^{-\lambda t}\, \mathbb{E}[e^{\lambda X}] \;\leq\; \exp \underbrace{\Big(-\lambda t + \frac{\lambda^2 \nu^2}{2} \Big)}_{g(\lambda, t)}, \qquad \text{valid for all } \lambda \in [0, b^{-1}).$$

In order to complete the proof, it remains to compute, for each fixed $t \geq 0$, the quantity $g^*(t) := \inf_{\lambda \in [0, b^{-1})} g(\lambda, t)$. For each fixed $t > 0$, the unconstrained minimum of the function $g(\cdot,\, t)$ occurs at $\lambda^* = t/\nu^2$. If $0 \leq t < \frac{\nu^2}{b}$, then this unconstrained optimum corresponds to the constrained minimum as well, so that $g^*(t) = -\frac{t^2}{2\nu^2}$ over this interval.

Otherwise, we may assume that $t \geq \frac{\nu^2}{b}$. In this case, since the function $g(\cdot, t)$ is monotonically decreasing in the interval $[0, \lambda^*)$, the constrained minimum is achieved at the boundary point $\lambda^\dagger = b^{-1}$, and we have

$$g^*(t) \;=\; g(\lambda^\dagger, t) = -\frac{t}{b} + \frac{1}{2b}\,\frac{\nu^2}{b} \;\overset{(i)}{\leq}\; -\frac{t}{2b},$$

where inequality (i) uses the fact that $\frac{\nu^2}{b} \leq t$. □

As shown in Example 2.4, the sub-exponential property can be verified by explicitly computing or bounding the moment-generating function. This direct calculation may be impractical in many settings, so it is natural to seek alternative approaches. One such method is provided by control on the polynomial moments of X. Given a random variable X with mean $\mu = \mathbb{E}[X]$ and variance $\sigma^2 = \mathbb{E}[X^2] - \mu^2$, we say that *Bernstein's condition* with parameter b holds if

$$|\mathbb{E}[(X - \mu)^k]| \leq \frac{1}{2} k!\; \sigma^2\; b^{k-2} \qquad \text{for } k = 3, 4, \ldots. \tag{2.16}$$

One sufficient condition for Bernstein's condition to hold is that X be bounded; in particular, if $|X - \mu| \leq b$, then it is straightforward to verify that condition (2.16) holds. Even for bounded variables, our next result will show that the Bernstein condition can be used to obtain tail bounds that can be tighter than the Hoeffding bound. Moreover, Bernstein's condition is also satisfied by various unbounded variables, which gives it much broader applicability.

When X satisfies the Bernstein condition, then it is sub-exponential with parameters determined by σ^2 and b. Indeed, by the power series expansion of the exponential, we have

$$\begin{aligned}\mathbb{E}[e^{\lambda(X-\mu)}] &= 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty} \lambda^k \, \frac{\mathbb{E}[(X-\mu)^k]}{k!} \\ &\overset{(i)}{\leq} 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^2\sigma^2}{2} \sum_{k=3}^{\infty} \big(|\lambda|\, b\big)^{k-2}.\end{aligned}$$

where the inequality (i) makes use of the Bernstein condition (2.16). For any $|\lambda| < 1/b$, we can sum the geometric series to obtain

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq 1 + \frac{\lambda^2\sigma^2/2}{1-b|\lambda|} \overset{(ii)}{\leq} e^{\frac{\lambda^2\sigma^2/2}{1-b|\lambda|}}, \tag{2.17}$$

where inequality (ii) follows from the bound $1+t \leq \exp(t)$. Consequently, we conclude that

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

showing that X is sub-exponential with parameters $(\sqrt{2}\sigma, 2b)$.

As a consequence, an application of Proposition 2.2 directly leads to tail bounds on a random variable satisfying the Bernstein condition (2.16). However, the resulting tail bound can be sharpened slightly, at least in terms of constant factors, by making direct use of the upper bound (2.17). We summarize in the following:

Proposition 2.3 (Bernstein-type bound)**.** For any random variable satisfying the Bernstein condition (2.16), we have

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{\lambda^2\sigma^2/2}{1-b|\lambda|}} \qquad \text{for all } |\lambda| < \tfrac{1}{b}, \tag{2.18}$$

and moreover, the concentration inequality

$$\mathbb{P}[|X-\mu| \geq t] \leq 2e^{-\frac{t^2}{2\,(\sigma^2+bt)}} \qquad \text{for all } t \geq 0. \tag{2.19}$$

We proved inequality (2.18) in the discussion preceding this proposition. Using this bound on the MGF, the tail bound (2.19) follows by setting $\lambda = \frac{t}{bt+\sigma^2} \in [0, \frac{1}{b})$ in the Chernoff bound, and then simplifying the resulting expression.

Remark: Proposition 2.3 has an important consequence even for bounded random variables (i.e., those satisfying $|X-\mu| \leq b$). The most straightforward way to control such variables is by exploiting the boundedness to show that $(X - \mu)$ is sub-Gaussian with parameter b (see Exercise 2.4), and then applying a Hoeffding-type inequality (see Proposition 2.1). Alternatively, using the fact that any bounded variable satisfies the Bernstein condition (2.17), we can also apply Proposition 2.3, thereby obtaining the tail bound (2.19), that involves *both* the variance σ^2 and the bound b. This tail bound shows that for suitably small t, the variable X has sub-Gaussian behavior with parameter σ, as opposed to the parameter b that would arise from a Hoeffding approach. Since $\sigma^2 = \mathbb{E}[(X-\mu)^2] \leq b^2$, this bound is never worse; moreover, it is substantially better when $\sigma^2 \ll b^2$, as would be the case for a random variable that occasionally takes on large values, but has relatively small variance. Such variance-based control frequently plays a key role in obtaining optimal rates in statistical problems, as will be seen in later chapters. For bounded random variables, Bennett's inequality can be used to provide sharper control on the tails (see Exercise 2.7).

Like the sub-Gaussian property, the sub-exponential property is preserved under summation for independent random variables, and the parameters (σ, b) transform in a simple way. In particular, suppose that X_k, $k = 1, \ldots, n$ are independent, and that variable X_k is subexponential with parameters (ν_k, b_k), and has mean $\mu_k = \mathbb{E}[X_k]$. We compute the moment generating function

$$\mathbb{E}[e^{\lambda \sum_{k=1}^n (X_k - \mu_k)}] \stackrel{(i)}{=} \prod_{k=1}^n \mathbb{E}[e^{\lambda X_k}] \stackrel{(ii)}{\leq} \prod_{k=1}^n e^{\lambda^2 \nu_k^2/2}, \qquad \text{valid for all } |\lambda| < \big(\max_{k=1,\ldots,n} b_k\big)^{-1},$$

where the equality (i) follows from independence, and inequality (ii) follows since X_k is sub-exponential with parameters (ν_k, b_k). Thus, we conclude that the variable $\sum_{k=1}^n (X_k - \mu_k)$ is sub-exponential with the parameters (ν_*, b_*), where

$$b_* := \max_{k=1,\ldots,n} b_k, \quad \text{and} \quad \nu_* := \sqrt{\sum_{k=1}^n \nu_k^2/n}.$$

Using the same argument as in Proposition 2.2, this observation leads directly to the upper tail bound

$$\mathbb{P}\big[\frac{1}{n}\sum_{i=1}^n (X_k - \mu_k) \geq t\big] \leq \begin{cases} e^{-\frac{nt^2}{2\nu_*^2}} & \text{for } 0 \leq t \leq \frac{\nu_*^2}{b_*} \\ e^{-\frac{nt}{2b_*}} & \text{for } t > \frac{\nu_*^2}{b_*}. \end{cases} \tag{2.20}$$

along with similar two-sided tail bounds. Let us illustrate our development thus far with some examples.

Example 2.5 (χ^2-variables)**.** A chi-squared random variable with n degrees of freedom, denoted by $Y \sim \chi_n^2$, can be represented as the sum $Y = \sum_{k=1}^n Z_k^2$ where $Z_k \sim \mathcal{N}(0, 1)$

are i.i.d. variates. As discussed in Example 2.4, the variable Z_k^2 is sub-exponential with parameters $(2, 4)$. Consequently, since the variables $Z_k, k = 1, 2, \ldots, n$ are independent, the χ^2-variate Y is sub-exponential with parameters $(\sigma, b) = (2\sqrt{n}, 4)$, and the preceding discussion yields the two-sided tail bound

$$\mathbb{P}\Big[\big|\frac{1}{n}\sum_{k=1}^{n} Z_k^2 - 1\big| \geq t\Big] \leq 2e^{-nt^2/8}, \qquad \text{for all } t \in (0, 1). \tag{2.21}$$

♣

The concentration of χ^2-variables plays an important role in the analysis of procedures based on taking random projections. A classical instance of the random projection method is the Johnson-Lindenstrauss analysis of metric embedding.

Example 2.6 (Johnson-Lindenstrauss embedding)**.** As one application of χ^2-concentration, consider the following problem. Suppose that we are given m data points $\{u^1, \ldots, u^m\}$ lying in $\mathbb{R}^d$. If the data dimension d is large, then it might be too expensive to store the data set. This challenge motivates the design of a mapping $F : \mathbb{R}^d \to \mathbb{R}^n$ with $n \ll d$ that preserves some "essential" features of the data set, and then store only the projected data set $\{F(u^1), \ldots, F(u^m)\}$. For example, since many algorithms are based on pairwise distances, we might be interested in a mapping F with the guarantee that for all pairs (u^i, u^j), we have

$$(1-\delta)\,\|u^i - u^j\|_2^2 \;\leq \|F(u^i) - F(u^j)\|_2^2 \;\leq\; (1+\delta)\|u^i - u^j\|_2^2, \tag{2.22}$$

for some tolerance $\delta \in (0, 1)$. Of course, this is always possible if the projected dimension n is large enough, but the goal is to do it with relatively small n.

Constructing such a mapping that satisfies the condition (2.22) with high probability turns out to be straightforward as long as the projected dimension satisfies $n = \Omega(\frac{1}{\delta^2}\log m)$. Observe that the projected dimension is independent of the ambient dimension d, and scales only logarithmically with the number of data points m.

The construction is probabilistic: first form a random matrix $X \in \mathbb{R}^{n\times d}$ filled with independent $\mathcal{N}(0, 1)$ entries, which defines a linear mapping $F : \mathbb{R}^d \to \mathbb{R}^n$ via $u \mapsto Xu/\sqrt{n}$. We now verify that F satisfies condition (2.22) with high probability. Let $X_i \in \mathbb{R}^d$ denote the i^{th} row of X, and consider some fixed $u \neq 0$. Since X_i is a standard normal vector, the variable $\langle X_i,\, u/\|u\|_2\rangle$ follows a $\mathcal{N}(0, 1)$ distribution, and hence the quantity

$$Y \;:=\; \frac{\|Xu\|_2^2}{\|u\|_2^2} = \sum_{i=1}^{n}\langle X_i,\, u/\|u\|_2\rangle^2,$$

follows a chi-squared distribution with n degrees of freedom, using the independence of the rows. Therefore, applying the tail bound (2.21), we find that

$$\mathbb{P}\Big[\big|\frac{\|Xu\|_2^2}{n\,\|u\|_2^2} - 1\big| \geq \delta\Big] \leq 2e^{-n\delta^2/8} \qquad \text{for all } \delta \in (0,1).$$

Re-arranging and recalling the definition of F yields the bound

$$\mathbb{P}\Big[\frac{\|F(u)\|_2^2}{\|u\|_2^2} \notin [(1-\delta),(1+\delta)]\Big] \;\leq\; 2e^{-n\delta^2/8}, \qquad \text{for any fixed } 0 \neq u \in \mathbb{R}^d.$$

Noting that there are at most $\binom{m}{2}$ distinct pairs of data points, we apply the union bound to conclude that

$$\mathbb{P}\Big[\frac{\|F(u^i-u^j)\|_2^2}{\|u^i-u^j\|_2^2} \notin [(1-\delta),(1+\delta)] \quad \text{for some } u^i \neq u^j\Big] \leq 2\binom{m}{2} e^{-n\delta^2/8}.$$

For any $\epsilon \in (0,1)$ and $m \geq 2$, this probability can be driven below ϵ by choosing $n > \frac{16}{\delta^2}\log(m/\epsilon)$. ♣

As a useful summary, the following theorem provides several equivalent characterizations of sub-exponential variables:

Theorem 2.2 (Equivalent characterizations of sub-exponential variables)**.** For a zero-mean random variable X, the following statements are equivalent:

(I) There are non-negative numbers (ν, b) such that

$$\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\nu^2\lambda^2}{2}} \quad \text{for all } |\lambda| < \tfrac{1}{b}. \tag{2.23}$$

(II) There is a positive number $c_0 > 0$ such that $\mathbb{E}[e^{\lambda X}] < \infty$ for all $|\lambda| \leq c_0$.

(III) There are constants $c_1, c_2 > 0$ such that

$$\mathbb{P}[|X| \geq t] \leq c_1\, e^{-c_2\, t} \quad \text{for all } t > 0. \tag{2.24}$$

(IV) The quantity $\gamma := \sup_{k \geq 2} \Big[\frac{\mathbb{E}[X^k]}{k!}\Big]^{1/k}$ is finite.

See Appendix B for the proof of this claim.

■ 2.1.4 Some one-sided results

Up to this point, we have focused on two-sided forms of Bernstein's condition, which then yields to bounds on both the upper and lower tails. As we have seen, one sufficient condition for Bernstein's condition to hold is an absolute bound, say $|X| \leq b$ almost surely. Of course, if such a bound only holds in a one-sided way, it is still possible to derive one-sided bounds. In this section, we state and prove one such result.

Proposition 2.4 (One-sided Bernstein's inequality)**.** If $X \leq b$ almost surely, then

$$\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \leq \exp\Big(\frac{\frac{\lambda^2}{2}\mathbb{E}[X^2]}{1-\frac{b\lambda}{3}}\Big) \quad \text{for all } \lambda \in [0, 3/b). \tag{2.25}$$

Consequently, given n independent random variables with $X_i \leq b$ almost surely, we have

$$\mathbb{P}\big[\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \geq n\delta\big] \leq \exp\Big(-\frac{n\delta^2}{\frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i^2] + \frac{b\delta}{3}}\Big). \tag{2.26}$$

Of course, if a random variable is bounded from below, then the same result can be used to derive bounds on its lower tail; we simply apply the bound (2.26) to the random variable $-X$. In the special case of independent non-negative random variables $Y_i \geq 0$, we find that

$$\mathbb{P}\big[\sum_{i=1}^n (Y_i - \mathbb{E}[Y_i]) \leq n\delta\big] \leq \exp\Big(-\frac{n\delta^2}{\frac{2}{n}\sum_{i=1}^n \mathbb{E}[Y_i^2]}\Big). \tag{2.27}$$

Thus, we see that the lower tail of *any* non-negative random variable satisfies a bound of the sub-Gaussian type, albeit with the second moment instead of the variance.

The proof of Proposition 2.4 is quite straightforward given our development thus far.

Proof. Let us write $\mathbb{E}[e^{\lambda X}] = 1 + \lambda\mathbb{E}[X] + \frac{1}{2}\lambda^2\mathbb{E}[X^2 h(\lambda X)]$, where we have defined the function $h(u) = 2\frac{e^u - u - 1}{u^2} = 2\sum_{k=2}^\infty \frac{u^{k-2}}{k!}$. Observe that for all $x < 0$ and $x' \in [0, b]$ and $\lambda > 0$, we have

$$h(\lambda x) \;\leq\; h(0) \;\leq\; h(\lambda x') \leq\; h(\lambda b).$$

Consequently, since $X \leq b$ almost surely, we have $\mathbb{E}[X^2 h(\lambda X)] \leq \mathbb{E}[X^2]\, h(\lambda b)$, and hence

$$\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \leq 1 + \frac{1}{2}\,\lambda^2\mathbb{E}[X^2]h(\lambda b) \;\leq\; \exp\Big\{\frac{\lambda^2\mathbb{E}[X^2]}{2}\,h(\lambda b)\Big\}.$$

Consequently, the bound (2.25) will follow if we can show that $h(\lambda b) \leq (1-\frac{\lambda b}{3})^{-1}$ for $\lambda b < 3$. By applying the inequality $k! \geq 2\,(3^{k-2})$, valid for all $k \geq 2$, we find that

$$h(\lambda b) \;=\; 2\sum_{k=2}^{\infty}\frac{(\lambda b)^{k-2}}{k!} \leq \sum_{k=2}^{\infty}\big(\frac{\lambda b}{3}\big)^{k-2} \;=\; \frac{1}{1-\frac{\lambda b}{3}},$$

where the final step uses the condition $\lambda b \in [0,3)$.

In order to prove the upper tail bound (2.26), we apply the Chernoff bound, exploiting independence to apply the MGF bound (2.25) separately, and thereby find that

$$\mathbb{P}\big[\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \geq n\delta] \leq \exp\Big(-\lambda n\delta + \frac{\frac{\lambda^2}{2}\sum_{i=1}^{n}\mathbb{E}[X_i^2]}{1-\frac{b\lambda}{3}}\Big), \quad \text{valid for } b\lambda \in [0,3).$$

Substituting $\lambda = \frac{n\delta}{\sum_{i=1}^{n}\mathbb{E}[X_i^2]+\frac{n\delta b}{3}} \in [0, 3/b)$ and simplifying yields the bound. □

■ 2.2 Martingale-based methods

Up until this point, our techniques have provided various types of bounds on sums of independent random variables. Many problems require bounds on more general functions of random variables, and one classical approach is based on martingale decompositions. In this section, we describe some of the results in this area along with some examples. Our treatment is quite brief, so we refer the reader to the bibliographic section for additional references.

■ 2.2.1 Background

Let us begin by introducing a particular case of a martingale sequence that is especially relevant for obtaining tail bounds. Let $\{X_k\}_{k=1}^n$ be a sequence of independent random variables, and consider the random variable $f(X) = f(X_1 \ldots, X_n)$, for some function $f : \mathbb{R}^n \to \mathbb{R}$. Suppose that our goal is to obtain bounds the deviations of f from its mean. In order to do so, we consider the sequence of random variables given by $Y_0 = \mathbb{E}[f(X)]$, $Y_n = f(X)$, and

$$Y_k = \mathbb{E}[f(X) \mid X_1, \ldots, X_k] \qquad \text{for } k = 1, \ldots, n-1, \tag{2.28}$$

where we assume that all conditional expectations exist. Note that Y_0 is a constant, and the random variables Y_k will exhibit more fluctuations as we move along the sequence from Y_0 to Y_n. Based on the intuition, the martingale approach to tail bounds is based on the telescoping decomposition

$$Y_n - Y_0 \;=\; \sum_{k=1}^{n} \underbrace{\big(Y_k - Y_{k-1}\big)}_{D_k}$$

in which the deviation $f(X) - \mathbb{E}[f(X)]$ is written as a sum of increments $\{D_k\}_{k=1}^n$. As we will see, the sequence $\{Y_k\}_{k=1}^n$ is a particular example of a martingale sequence, known as the *Doob martingale*, whereas the sequence $\{D_k\}_{k=1}^n$ is an example of a martingale difference sequence.

With this example in mind, we now turn to the general definition of a martingale sequence. Let $\{\mathcal{F}_k\}_{k=1}^\infty$ be a sequence of σ-fields that are nested, meaning that $\mathcal{F}_k \subseteq \mathcal{F}_{k+1}$ for all $k \geq 1$; such a sequence is known as a filtration. In the Doob martingale described above, the σ-field $\sigma(X_1, \ldots, X_k)$ generated by the first k variables plays the role of $\mathcal{F}_k$. Let $\{Y_k\}_{k=1}^\infty$ be a sequence of random variables such that Y_k is measurable with respect to the σ-field $\mathcal{F}_k$. In this case, we say that $\{Y_k\}_{k=1}^\infty$ is adapted to the filtration $\{\mathcal{F}_k\}_{k=1}^\infty$. In the Doob martingale, the random variable Y_k is a measurable function of $(X_1, \ldots, X_k)$, and hence the sequence is adapted to the filtration defined by the σ-fields. We are now ready to define a general martingale:

Definition 2.3. Given a sequence $\{Y_k\}_{k=1}^\infty$ of random variables adapted to a filtration $\{\mathcal{F}_k\}_{k=1}^\infty$, the pair $\{(Y_k, \mathcal{F}_k)\}_{k=1}^\infty$ is a *martingale* if for all $k \geq 1$,

$$\mathbb{E}[|Y_k|] < \infty, \quad \text{and} \quad \mathbb{E}\big[Y_{k+1} \mid \mathcal{F}_k\big] = Y_k. \tag{2.29}$$

It is frequently the case that the filtration is defined by a second sequence of random variables $\{X_k\}_{k=1}^\infty$ via the canonical σ-fields $\mathcal{F}_k := \sigma(X_1, \ldots, X_k)$. In this case, we say that $\{Y_k\}_{k=1}^\infty$ is a martingale sequence with respect to $\{X_k\}_{k=1}^\infty$. The Doob construction is an instance of such a martingale sequence. If a sequence is martingale with respect to itself (i.e., with $\mathcal{F}_k = \sigma(Y_1, \ldots, Y_k)$), then we say simply that $\{Y_k\}_{k=1}^\infty$ forms a martingale sequence.

Let us consider some examples to illustrate:

Example 2.7 (Partial sums as martingales)**.** Perhaps the simplest instance of a martingale is provided by considering partial sums of an i.i.d. sequence. Let $\{X_k\}_{k=1}^\infty$ be a sequence of i.i.d. random variables with mean μ, and define the partial sums $S_k := \sum_{j=1}^k X_j$. Defining $\mathcal{F}_k = \sigma(X_1, \ldots, X_k)$, the random variable S_k is measurable with respect to $\mathcal{F}_k$, and moreover, we have

$$\begin{aligned}\mathbb{E}[S_{k+1} \mid \mathcal{F}_k] &= \mathbb{E}[X_{k+1} + S_k \mid X_1, \ldots X_k] \\ &= \mathbb{E}[X_{k+1}] + S_k \;=\; \mu + S_k.\end{aligned}$$

Here we have used the facts that X_{k+1} is independent of $X_1^k := (X_1, \ldots, X_k)$, and that S_k is a function of X_1^k. Thus, while the sequence $\{S_k\}_{k=1}^\infty$ itself is not a martingale unless $\mu = 0$, the recentered variables $Y_k := S_k - k\mu$ for $k \geq 1$ define a martingale sequence with respect to $\{X_k\}_{k=1}^\infty$. $\clubsuit$

Let us now show that the Doob construction does lead to a martingale, as long as the underlying function f is absolutely integrable.

Example 2.8 (Doob construction)**.** Recall the sequence $Y_k = \mathbb{E}[f(X) \mid X_1, \ldots, X_k]$ previously defined, and suppose that $\mathbb{E}[|f(X)|] < \infty$. We claim that $\{Y_k\}_{k=0}^n$ is a martingale with respect to $\{X_k\}_{k=1}^n$. Indeed, using the shorthand $X_1^k = (X_1, X_2, \ldots, X_k)$, we have

$$\mathbb{E}[|Y_k|] = \mathbb{E}\big[\big|\mathbb{E}[f(X) \mid X_1^k]\big] \;\leq\; \mathbb{E}[|f(X)|] < \infty,$$

where the bound follows from Jensen's inequality. Turning to the second property, we have

$$\mathbb{E}[Y_{k+1} \mid X_1^k] = \mathbb{E}\Big[\mathbb{E}[f(X) \mid X_1^{k+1}] \mid X_1^k\Big] \stackrel{(i)}{=} \mathbb{E}[f(X) \mid X_1^k] \;=\; Y_k,$$

where we have used the tower property of conditional expectation in step (i). $\clubsuit$

The following martingale plays an important role in analyzing stopping rules for sequential hypothesis tests:

Example 2.9 (Likelihood ratio)**.** Let f and g be two mutually continuous densities, and let $\{X_k\}_{k=1}^\infty$ be a sequence of random variables drawn i.i.d. according to f. If we let $Y_n := \prod_{k=1}^n g(X_k)/f(X_k)$ be the likelihood ratio based on the first n samples, then sequence $\{Y_k\}_{k=1}^\infty$ is a martingale with respect to $\{X_k\}_{k=1}^\infty$. Indeed, we have

$$\mathbb{E}[Y_{n+1} \mid X_1, \ldots, X_n] = \mathbb{E}\Big[\frac{g(X_{n+1})}{f(X_{n+1})}\Big] \prod_{k=1}^n \frac{g(X_k)}{f(X_k)} \;=\; Y_n,$$

using the fact that $\mathbb{E}[\frac{g(X_{n+1})}{f(X_{n+1})}] = 1$. $\clubsuit$

A closely related notion is that of *martingale difference sequence*, meaning an adapted sequence $\{(D_k, \mathcal{F}_k)\}_{k=1}^\infty$ such that for all $k \geq 1$,

$$\mathbb{E}[|D_k|] < \infty, \quad \text{and} \quad \mathbb{E}[D_{k+1} \mid \mathcal{F}_k] = 0. \tag{2.30}$$

As suggested by their name, such difference sequences arise in a natural way from martingales. In particular, given a martingale $\{(Y_k, \mathcal{F}_k)\}_{k=0}^{\infty}$, let us define $D_k = Y_k - Y_{k-1}$ for $k \geq 1$. We then have

$$\begin{aligned}
\mathbb{E}[D_{k+1} \mid \mathcal{F}_k] &= \mathbb{E}[Y_{k+1} \mid \mathcal{F}_k] - \mathbb{E}[Y_k \mid \mathcal{F}_k] \\
&= \mathbb{E}[Y_{k+1} \mid \mathcal{F}_k] - Y_k \;=\; 0,
\end{aligned}$$

using the martingale property (2.29) and the fact that Y_k is measurable with respect to $\mathcal{F}_k$. Thus, for any martingale sequence $\{Y_k\}_{k=0}^{\infty}$, we have the telescoping decomposition $Y_n - Y_0 = \sum_{k=1}^n D_k$ where $\{D_k\}_{k=1}^{\infty}$ is martingale difference sequence. This decomposition plays an important role in our development of concentration inequalities to follow.

2.2.2 Concentration bounds for martingale difference sequences

We now turn to the derivation of concentration inequalities for martingales. These inequalities can be viewed in one of two ways: either as bounds for the difference $Y_n - Y_0$, or as bounds for the sum $\sum_{k=1}^n D_k$ of the associated martingale difference sequence (MDS). Throughout this section, we present results mainly in terms of martingale differences, with the understanding that such bounds have direct consequences for martingale sequences. Of particular interest to us is the Doob martingale described in Example 2.8, which can be used to control the deviations of a function from its expectation.

We begin by stating and proving a general Bernstein-type bound for a MDS, based on imposing a sub-exponential condition on the martingale differences.

Theorem 2.3. Let $\{(D_k, \mathcal{F}_k)\}_{k=1}^{\infty}$ be a martingale difference sequence, and suppose that for any $|\lambda| < 1/b_k$, we have $\mathbb{E}[e^{\lambda D_k} \mid \mathcal{F}_{k-1}] \leq e^{\lambda^2 \nu_k^2/2}$ almost surely. Then

(a) The sum $\sum_{k=1}^n D_k$ is sub-exponential with parameters $(\sqrt{\sum_{k=1}^n \nu_k^2},\; b_*)$ where $b_* := \max_{k=1,\ldots,n} b_k$.

(b) Consequently, for all $t \geq 0$,

$$\mathbb{P}\big[|\sum_{k=1}^n D_k| \geq t\big] \leq \begin{cases} 2\, e^{-\frac{t^2}{2\sum_{k=1}^n \nu_k^2}} & \text{if } 0 \leq t \leq \frac{\sum_{k=1}^n \nu_k^2}{b_*}, \quad \text{and} \\ 2\, e^{-\frac{t}{2\, b_*}} & \text{if } t > \frac{\sum_{k=1}^n \nu_k^2}{b_*}. \end{cases} \tag{2.31}$$

Proof. We follow the standard approach of controlling the moment generating function of $\sum_{k=1}^n D_k$, and then applying the Chernoff bound. Letting $\lambda \in (-1/b_*, 1/b_*)$ be arbitrary, conditioning on $\mathcal{F}_{n-1}$ and applying iterated expectation yields

$$\begin{aligned}\mathbb{E}\big[e^{\lambda\left(\sum_{k=1}^{n} D_k\right)}\big] &= \mathbb{E}\Big[e^{\lambda\left(\sum_{k=1}^{n-1} D_k\right)} \;\; \mathbb{E}[e^{\lambda D_n} \mid \mathcal{F}_{n-1}]\Big] \\ &\leq \mathbb{E}\big[e^{\lambda \sum_{k=1}^{n-1} D_k}\big] \; e^{\lambda^2 \nu_n^2/2}, \end{aligned} \tag{2.32}$$

where the inequality follows from the stated assumption on D_n. Iterating this procedure yields the bound $\mathbb{E}\big[e^{\lambda \sum_{k=1}^{n} D_k}\big] \leq e^{\lambda^2 \sum_{k=1}^{n} \nu_k^2/2}$, valid for all $|\lambda| < \frac{1}{b_*}$. By definition, we conclude that $\sum_{k=1}^{n} D_k$ is sub-exponential with parameters $(\sqrt{\sum_{k=1}^{n} \nu_k^2}, b_*)$, as claimed. The tail bound (2.31) follows by applying Proposition 2.2. □

In order for Theorem 2.3 to be useful in practice, we need to isolate sufficient and easily checkable conditions for the differences D_k to be almost surely sub-exponential (or sub-Gaussian when $b = +\infty$). As discussed previously, bounded random variables are sub-Gaussian, which leads to the following corollary:

Corollary 2.1 (Azuma-Hoeffding)**.** Let $(\{(D_k, \mathcal{F}_k)\}_{k=1}^{\infty})$ be a martingale difference sequence, and suppose that $|D_k| \leq b_k$ almost surely for all $k \geq 1$. Then for all $t \geq 0$,

$$\mathbb{P}\Big[|\sum_{k=1}^{n} D_k| \geq t\Big] \leq 2e^{-\frac{2t^2}{\sum_{k=1}^{n} b_k^2}}. \tag{2.33}$$

Proof. Recall the decomposition (2.32) in the proof of Theorem 2.3; from the structure of this argument, it suffices to show that $\mathbb{E}[e^{\lambda D_k} \mid \mathcal{F}_{k-1}] \leq e^{\lambda^2 b_k^2/2}$ a.s. for each $k = 1, 2, \ldots, n$. But since $|D_k| \leq b_k$ almost surely, the conditioned variable $(D_k \mid \mathcal{F}_{k-1})$ is also bounded almost surely, and hence from the result of Exercise 2.4, it is sub-Gaussian with parameter at most $\sigma = b_k$. □

An important application of Corollary 2.1 concerns functions that satisfy a bounded difference property. In particular, we say that $f : \mathbb{R}^n \rightarrow \mathbb{R}$ satisfies the bounded difference inequality with parameters $(L_1, \ldots, L_n)$ if for each $k = 1, 2, \ldots, n$,

$$|f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{k-1}, x_k', x_{k+1}, \ldots, x_n)| \leq L_k \qquad \text{for all } x, x' \in \mathbb{R}^n. \tag{2.34}$$

For instance, if the function f is L-Lipschitz with respect to the Hamming norm $d_H(x, y) = \sum_{i=1}^{n} \mathbb{I}[x_i \neq y_i]$, which counts the number of positions in which x and y differ, then the bounded difference inequality holds with parameter L uniformly across all co-ordinates.

Corollary 2.2 (Bounded differences inequality)**.** Suppose that f satisfies the bounded difference property (2.34) with parameters $(L_1, \ldots, L_n)$ and that the random vector $X = (X_1, X_2, \ldots, X_n)$ has independent components. Then 1

$$\mathbb{P}\big[|f(X) - \mathbb{E}[f(X)]| \geq t\big] \leq 2e^{-\frac{2t^2}{\sum_{k=1}^n L_k^2}} \quad \text{for all } t \geq 0. \tag{2.35}$$

2

3

Proof. Recalling the Doob martingale introduced in Example 2.8, consider the associated martingale difference sequence

$$D_k = \mathbb{E}[f(X) \mid X_1, \ldots, X_k] - \mathbb{E}[f(X) \mid X_1, \ldots, X_{k-1}]. \tag{2.36}$$

If we define the function $g : \mathbb{R}^k \to \mathbb{R}$ via $g(x_1, \ldots, x_k) := \mathbb{E}[f(X) \mid x_1, \ldots, x_k]$, the independence assumption means that we can write

$$D_k = g(X_1, \ldots, X_k) - \mathbb{E}_{X_k'}\big[g(X_1, \ldots, X_k')\big] \;=\; \mathbb{E}_{X_k'}\big[g(X_1, \ldots, X_k) - g(X_1, \ldots, X_k')\big],$$

where X_k' is an independent copy of X_k. Thus, if we can show that g satisfies the bounded difference property with parameter L_k in co-ordinate k, then Corollary 2.1 yields the claim. But by independence and the definition of g, for any pair of k-tuples $(x_1, \ldots, x_k)$ and $(x_1, \ldots, x_k')$, we can write

$$\begin{aligned} g(x_1, \ldots, x_k) - g(x_1, \ldots, x_k') &= \mathbb{E}_{X_{k+1}^n}\big[f(x_1, \ldots, x_k, X_{k+1}, \ldots, X_n) - f(x_1, \ldots, x_k', X_{k+1}, \ldots, X_n)\big] \\ &\leq L_k, \end{aligned}$$

using the bounded differences inequality for f. $\square$ 4

Remark: In the special case when f is L-Lipschitz with respect to the Hamming norm, Corollary 2.2 implies that

$$\mathbb{P}\big[|f(X) - \mathbb{E}[f(X)]| \geq t\big] \leq 2e^{-\frac{2t^2}{nL^2}} \quad \text{for all } t \geq 0. \tag{2.37}$$

Let us consider some examples to illustrate. 5

Example 2.10 (U-statistics)**.** Let $g : \mathbb{R}^2 \to \mathbb{R}$ be a symmetric function of its arguments. Given an i.i.d. sequence $X_k, k \geq 1$ of random variables, the quantity

$$U := \frac{1}{\binom{n}{2}} \sum_{j \neq k} g(X_j, X_k) \tag{2.38}$$

is known as a pairwise *U-statistic*. For instance, if $g(s,t) = |s-t|$, then U is an unbiased estimator of the mean absolute deviation $\mathbb{E}[|X_1 - X_2|]$. Note that while U is not a sum of independent random variables, the dependence is relatively weak, and this is revealed by a martingale analysis. If g is bounded (say $\|g\|_\infty \leq b$), then Corollary 2.2 can be used to establish concentration of U around its mean. Viewing U as a function $f(X_1, \ldots, X_n)$, for any given co-ordinate k, we have

$$\begin{aligned}
|f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{k-1}, x_k', x_{k+1}, \ldots, x_n)| &\leq \frac{1}{\binom{n}{2}} \sum_{j \neq k} |g(x_j, x_k) - g(x_j, x_k')| \\
&\leq \frac{(n-1)\,(2b)}{\binom{n}{2}} \;=\; \frac{4b}{n},
\end{aligned}$$

so that the bounded differences property holds with parameter $L_k = \frac{4b}{n}$ in each co-ordinate. Thus, we conclude that

$$\mathbb{P}\big[|U - \mathbb{E}[U]| \geq t\big] \leq 2e^{-\frac{nt^2}{8b^2}}.$$

This tail inequality implies that U is a consistent estimate of $\mathbb{E}[U]$, and also yields finite sample bounds on its quality as an estimator. Similar techniques can be used to obtain tail bounds on U-statistics of higher order, involving sums over k tuples of variables. ♣

Example 2.11 (Rademacher complexity)**.** Let $\varepsilon_k, k \geq 1$ be an i.i.d. sequence of Rademacher variables (i.e., taking the values $\{-1, +1\}$ equiprobably, as in Example 2.2). Given a collection of vectors $\mathcal{A} \subset \mathbb{R}^n$, define the random variable[1]

$$Z := \sup_{a \in \mathcal{A}} \Big[\sum_{k=1}^n a_k \varepsilon_k\Big] \;=\; \sup_{a \in \mathcal{A}} \big[\langle a,\, \varepsilon\rangle\big]. \tag{2.39}$$

The random variable Z measures the size of $\mathcal{A}$ in a certain sense, and its expectation $\mathcal{R}(\mathcal{A}) := \mathbb{E}[Z(\mathcal{A})]$ is known as the *Rademacher complexity* of the set $\mathcal{A}$.

Corollary 2.2 can be used to show that $Z(\mathcal{A})$ is sub-Gaussian. Viewing $Z(\mathcal{A})$ as a function $f(\varepsilon_1, \ldots, \varepsilon_n)$, we need to bound the maximum change when co-ordinate k is changed. Let ε' be an n-vector with $\varepsilon_j' = \varepsilon_j$ for all $j \neq k$, and let $a \in \mathcal{A}$ be arbitrary. Since $f(\varepsilon') \geq \langle a,\, \varepsilon'\rangle$ for any $a \in \mathcal{A}$, we have

$$\langle a,\, \varepsilon\rangle - f(\varepsilon') \leq \langle a,\, \varepsilon - \varepsilon'\rangle = a_k\,(\varepsilon_k - \varepsilon_k') \;\leq\; 2|a_k|.$$

Taking the supremum over $\mathcal{A}$ on both sides, we obtain the inequality

$$f(\varepsilon) - f(\varepsilon') \leq 2 \sup_{a \in \mathcal{A}} |a_k|.$$

[1]For the reader concerned about measurability, see the bibliographic discussion in Chapter 4

Since the same argument applies with the roles of ε and ε' reversed, we conclude that f satisfies the bounded difference inequality in co-ordinate k with parameter $2\sup_{a\in\mathcal{A}} |a_k|$. Consequently, Corollary 2.2 implies that the the random variable $Z(\mathcal{A})$ is sub-Gaussian with parameter at most $4\sum_{k=1}^n \sup_{a\in\mathcal{A}} a_k^2$. This sub-Gaussian parameter can be reduced to $4\sup_{a\in\mathcal{A}} \sum_{k=1}^n a_k^2$ using alternative techniques (see Example 3.2 in Chapter 3). ♣

■ 2.3 Lipschitz functions of Gaussian variables

We conclude this chapter with a classical result on the concentration properties of Lipschitz functions of Gaussian variables. These functions exhibit a particularly attractive form of dimension-free concentration. Let us say that function $f : \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz with respect to the Euclidean norm $\|\cdot\|_2$ if

$$|f(x) - f(y)| \leq L\,\|x - y\|_2 \quad \text{for all } x, y \in \mathbb{R}^n. \tag{2.40}$$

The following result guarantees that any such function is sub-Gaussian with parameter at most L:

Theorem 2.4. Let $(X_1, \ldots, X_n)$ be a vector of i.i.d. standard Gaussian variables, and let $f : \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz with respect to the Euclidean norm. Then the variable $f(X) - \mathbb{E}[f(X)]$ is sub-Gaussian with parameter at most L, and hence

$$\mathbb{P}\Big[\big|f(X) - \mathbb{E}[f(X)]\big| \geq t\Big] \leq 2e^{-\frac{t^2}{2L^2}} \quad \text{for all } t \geq 0. \tag{2.41}$$

Note that this result is truly remarkable: it guarantees that any L-Lipschitz function of a standard Gaussian random vector, regardless of the dimension, exhibits concentration like a scalar Gaussian variable with variance L^2.

Proof. With the aim of keeping the proof as simple as possible, let us prove a version of the concentration bound (2.41) with a weaker constant in the exponent. (See the bibliographic notes for references to proofs of the sharpest results.) We also prove the result for a function that is both Lipschitz *and* differentiable; since any Lipschitz function is differentiable almost everywhere[2], it is then straightforward to extend this result to the general setting. In order to prove this version of the theorem, we begin by stating an auxiliary technical lemma:

[2]This fact is a consequence of Rademacher's theorem.

Lemma 2.1. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable. Then for any convex function $\phi : \mathbb{R} \to \mathbb{R}$, we have

1
$$\mathbb{E}\Big[\phi\big(f(X) - \mathbb{E}[f(X)]\big)\Big] \leq \mathbb{E}\Big[\phi\big(\frac{\pi}{2}\langle \nabla f(X),\, Y\rangle\big)\Big] \tag{2.42}$$

where $X, Y \sim \mathcal{N}(0, \mathbf{I}_n)$ are standard multivariate Gaussian, and independent.

2

We now prove the theorem using this lemma. For any fixed $\lambda \in \mathbb{R}$, applying inequality (2.42) to the convex function $t \mapsto \exp(\lambda t)$ yields

$$\begin{aligned}
\mathbb{E}\Big[\exp\big(\lambda\big\{f(X) - \mathbb{E}[f(X)]\big\}\big)\Big] &\leq \mathbb{E}\Big[\exp\big(\frac{\lambda\,\pi}{2}\sum_{k=1}^n Y_k\, \frac{\partial f}{\partial x_k}(X)\big)\Big] \\
&= \mathbb{E}_X\Big[\mathbb{E}_Y\Big[\exp\big(\frac{\lambda\,\pi}{2}\sum_{k=1}^n Y_k\, \frac{\partial f}{\partial x_k}(X)\big)\Big]\Big] \\
&= \mathbb{E}_X\Big[\prod_{k=1}^n \mathbb{E}_{Y_k}\big[\exp\big(\frac{\lambda\,\pi}{2} Y_k\, \frac{\partial f}{\partial x_k}(X)\big)\big]\Big],
\end{aligned}$$

where we have used the independence of X and Y, and the i.i.d. nature of the components of Y. Since $Y_k \sim \mathcal{N}(0,1)$, we have

$$\mathbb{E}_{Y_k}\big[\exp\big(\frac{\lambda\,\pi}{2} Y_k\, \frac{\partial f}{\partial x_k}(X)\big)\big] = \exp\Big(\frac{\lambda^2\,\pi^2}{8}\big(\frac{\partial f}{\partial x_k}(X)\big)^2\Big),$$

and hence

$$\mathbb{E}\big[e^{\lambda\big\{f(X) - \mathbb{E}[f(X)]\big\}}\big] \leq \mathbb{E}\big[e^{\frac{\lambda^2\pi^2}{8}\|\nabla f(X)\|_2^2}\big] \;\leq\; e^{\frac{1}{8}\,\lambda^2\pi^2\,L^2},$$

where the final inequality follows from the fact that $\|\nabla f(X)\|_2 \leq L$, due to the Lipschitz condition on f. We have thus shown that $f(X) - \mathbb{E}[f(X)]$ is sub-Gaussian with parameter at most $\frac{\pi L}{2}$, from which the tail bound

$$\mathbb{P}\Big[\big|f(X) - \mathbb{E}[f(X)]\big| \geq t\Big] \leq 2\exp\Big(-\frac{2\,t^2}{\pi^2 L^2}\Big) \quad \text{for all } t \geq 0$$

3 follows from Proposition 2.1.

4

It remains to prove Lemma 2.1, and we do so via an interpolation method that exploits the rotation invariance of the Gaussian distribution. For each $\theta \in [0, \pi/2]$, consider the random vector $Z(\theta) \in \mathbb{R}^n$ with components

$$Z_k(\theta) := X_k \sin\theta + Y_k \cos\theta \quad \text{for } k = 1, 2, \ldots, n.$$

By the rotation invariance of the Gaussian distribution, we observe that for all $\theta \in [0, \pi/2]$, the pair $(Z_k(\theta), Z'_k(\theta))$ is a jointly Gaussian vector, with zero mean and identity covariance $\mathbf{I}_2$. By convexity of ϕ, we have

$$\mathbb{E}_X\big[\phi\big(f(X) - \mathbb{E}_Y[f(Y)]\big)\big] \leq \mathbb{E}_{X,Y}\big[\phi\big(f(X) - f(Y)\big)\big]. \tag{2.43}$$

Now since $Z_k(0) = Y_k$ and $Z_k(\pi/2) = X_k$ for all $k = 1, \ldots, n$, we have

$$f(X) - f(Y) = \int_0^{\pi/2} \frac{d}{d\theta} f(Z(\theta)) d\theta \;=\; \int_0^{\pi/2} \langle \nabla f(Z(\theta)),\, Z'(\theta)\rangle d\theta,$$

and hence

$$\begin{aligned}
\mathbb{E}_X\big[\phi\big(f(X) - \mathbb{E}_Y[f(Y)]\big)\big] &\leq \mathbb{E}_{X,Y}\big[\phi\big(\int_0^{\pi/2} \langle \nabla f(Z(\theta)),\, Z'(\theta)\rangle d\theta\big)\big] \\
&= \mathbb{E}_{X,Y}\big[\phi\big(\frac{1}{\pi/2}\int_0^{\pi/2} \frac{\pi}{2}\langle \nabla f(Z(\theta)),\, Z'(\theta)\rangle d\theta\big)\big] \\
&\leq \frac{1}{\pi/2}\int_0^{\pi/2} \mathbb{E}_{X,Y}\big[\phi\big(\frac{\pi}{2}\langle \nabla f(Z(\theta)),\, Z'(\theta)\rangle\big)\big] d\theta,
\end{aligned}$$

where the final step again uses convexity of ϕ. But since $(Z_k(\theta), Z'_k(\theta)) \sim \mathcal{N}(0, \mathbf{I}_2)$ for all θ, the expectation does not depend on θ, and so we have

$$\frac{1}{\pi/2}\int_0^{\pi/2} \mathbb{E}_{X,Y}\big[\phi\big(\frac{\pi}{2}\langle \nabla f(Z(\theta)),\, Z'(\theta)\rangle\big)\big] d\theta = \mathbb{E}\big[\phi\big(\frac{\pi}{2}\langle \nabla f(\widetilde{X}),\, \widetilde{Y}\rangle\big)\big]$$

where $(\widetilde{X}, \widetilde{Y})$ are independent standard Gaussian n-vectors. This completes the proof of the bound (2.42).

□

Note that the proof makes essential use of various properties specific to the standard Gaussian distribution. However, similar concentration results hold for other non-Gaussian distributions, including the uniform distribution on the sphere and any strictly log-concave distribution (see Chapter 3 for further discussion of such distributions). Without additional structure of the function f (such as convexity), dimension-free concentration for Lipschitz functions need not hold for an arbitrary sub-Gaussian distribution; see the bibliographic section for further discussion of this fact.

Theorem 2.4 is useful for a broad range of problems; let us consider some examples to illustrate.

Example 2.12 (χ^2 concentration)**.** If $Z_k \sim \mathcal{N}(0, 1)$ are i.i.d. standard normal variates, then $Y := \sum_{k=1}^n Z_k^2$ is a chi-squared variate with n degrees of freedom. The most direct way to obtain tail bounds for Y is by noting that Z_k^2 is sub-exponential, and exploiting independence (see Example 2.5). In this example, we pursue an alternative approach—namely, via concentration for Lipschitz functions of Gaussian variates. Indeed, defining the variable $V = \sqrt{Y}/\sqrt{n}$, we can write $V = \|(Z_1, \ldots, Z_n)\|_2/\sqrt{n}$, and since the Euclidean norm is a 1-Lipschitz function, Theorem 2.4 implies that

$$\mathbb{P}[V \geq \mathbb{E}[V] + \delta] \leq \exp(-n\delta^2/2) \quad \text{for all } \delta \geq 0.$$

Using concavity of the square root function and Jensen's inequality, we have

$$\mathbb{E}[V] \leq \sqrt{\mathbb{E}[V^2]} \;=\; \Big\{\frac{1}{n}\sum_{i=1}^n \mathbb{E}[Z_k^2]\Big\}^{1/2} \;=\; 1.$$

Recalling that $V = \sqrt{Y}/\sqrt{n}$ and putting together the pieces yields

$$\mathbb{P}[\frac{Y}{n} \geq (1+\delta)^2] \leq \exp(-n\delta^2/2) \quad \text{for all } \delta \geq 0.$$

Since $(1+\delta)^2 = 1 + 2\delta + \delta^2 \leq 1 + 3\delta$ for all $\delta \in [0,1]$, we conclude that

$$\mathbb{P}\big[Y \geq n\,(1+t)\big] \leq \exp(-nt^2/18) \quad \text{for all } t \in [0,3], \tag{2.44}$$

where we have made the substitution $t = 3\delta$. It is worthwhile comparing this tail bound to those that can be obtained by using the fact that each Z_k^2 is sub-exponential, as discussed in Example 2.5. ♣

Example 2.13 (Order statistics)**.** Given a random vector $(X_1, X_2, \ldots, X_n)$, its order statistics are obtained by re-ordering the samples in a non-increasing manner—namely as

$$X_{(1)} \geq X_{(2)} \geq \ldots \geq X_{(n-1)} \geq X_{(n)}. \tag{2.45}$$

As particular cases, we have $X_{(1)} = \max_{k=1,\ldots,n} X_k$ and $X_{(n)} = \min_{k=1,\ldots,n} X_k$. Given another random vector $(Y_1, \ldots, Y_n)$, it can be shown that $|X_{(k)} - Y_{(k)}| \leq \|X - Y\|_2$ for all $k = 1, \ldots, n$, so that each order statistic is a 1-Lipschitz function. (We leave the verification of this inequality as an exercise for the reader.) Consequently, when X is a Gaussian random vector, Theorem 2.4 implies that

$$\mathbb{P}[|X_{(k)} - \mathbb{E}[X_{(k)}]| \geq \delta] \leq 2e^{-\frac{\delta^2}{2}} \quad \text{for all } \delta \geq 0.$$

♣

Example 2.14 (Gaussian complexity)**.** This example is closely related to our earlier discussion of Rademacher complexity in Example 2.11. Let $\{w_k\}_{k=1}^n$ be an i.i.d. sequence of $\mathcal{N}(0,1)$ variables. Given a collection of vectors $\mathcal{A} \subset \mathbb{R}^n$, define the random variable[3]

$$Z := \sup_{a \in \mathcal{A}} \Big[\sum_{k=1}^{n} a_k w_k \Big] \;=\; \sup_{a \in \mathcal{A}} \langle a,\, w \rangle. \tag{2.46}$$

As with the Rademacher complexity, the variable $Z = Z(\mathcal{A})$ is one way of measuring the size of the set $\mathcal{A}$, and will play an important role in later chapters. Viewing Z as a function $f(w_1, \ldots, w_n)$, let us verify that f is Lipschitz (with respect to Euclidean norm) with parameter $\sup_{a \in \mathcal{A}} \|a\|_2$. Let $w, w' \in \mathbb{R}^n$ be arbitrary, and let $a^* \in \mathcal{A}$ be any vector that achieves the maximum defining $f(w)$. Following the same argument as Example 2.11, we have the upper bound

$$f(w) - f(w') \leq \langle a^*,\, w - w' \rangle \;\leq\; \mathcal{W}(\mathcal{A})\, \|w - w'\|_2.$$

where $\mathcal{W}(\mathcal{A}) = \sup_{a \in \mathcal{A}} \|a\|_2$ is the Euclidean width of the set. The same argument holds with the roles of w and w' reversed, and hence $|f(w) - f(w')| \leq \mathcal{W}(\mathcal{A})\, \|w - w'\|_2$. Consequently, Theorem 2.4 implies that

$$\mathbb{P}\big[|Z - \mathbb{E}[Z]| \geq \delta\big] \leq 2 \exp\big(-\frac{\delta^2}{2\mathcal{W}^2(\mathcal{A})}\big). \tag{2.47}$$

♣ 1

Example 2.15 (Gaussian chaos variables)**.** As a generalization of the previous example, let $\mathbf{Q} \in \mathbb{R}^{n \times n}$ be a symmetric matrix, and let $w, \widetilde{w}$ be independent zero-mean Gaussian random vectors with covariance matrix $\mathbf{I}_n$. The random variable

$$Z := \sum_{i,j=1}^{n} Q_{ij} w_i \widetilde{w}_j$$

is known as a (decoupled) Gaussian chaos. By the independence of w and $\widetilde{w}$, we have 2
$\mathbb{E}[Z] = 0$, so it is natural to seek on a tail bound on the event $\{|Z| \geq t\}$. 3

Conditioned on $\widetilde{w}$, the variable $(Z \mid \widetilde{w})$ is a zero-mean Gaussian variable with variance $\widetilde{w}^T \mathbf{Q}^2 \widetilde{w}$, and consequently, we have the tail bound

$$\mathbb{P}[|Z| \geq \delta \mid \widetilde{w}] \leq 2 e^{-\frac{\delta^2}{2 \widetilde{w}^T \mathbf{Q}^2 \widetilde{w}}}. \tag{2.48}$$

Let us now bound the random variable $Y = \|\mathbf{Q}\widetilde{w}\|_2$. It is a Lipschitz function of the Gaussian vector $\widetilde{w}$, with Lipschitz constant given by the operator norm $|\!|\!|\mathbf{Q}|\!|\!|_{\mathrm{op}} = \sup_{\|u\|_2 = 1} \|\mathbf{Q}u\|_2$. Moreover, by Jensen's inequality, we have $\mathbb{E}[Y] \leq \sqrt{\mathbb{E}[\widetilde{w}^T \mathbf{Q}^2 \widetilde{w}]} = |\!|\!|\mathbf{Q}|\!|\!|_{\mathrm{F}}$. Putting to-

[3]For measurability concerns, see the bibliographic discussion in Chapter 4.

gether the pieces yields the tail bound

$$\mathbb{P}\big[\|\mathbf{Q}\widetilde{w}\|_2 \geq |\!|\!|\mathbf{Q}|\!|\!|_{\mathrm{F}} + t\big] \leq 2\exp\big(-\frac{t^2}{2|\!|\!|\mathbf{Q}|\!|\!|_{\mathrm{op}}^2}\big).$$

Setting $t^2 = 4\delta|\!|\!|\mathbf{Q}|\!|\!|_{\mathrm{op}}$ and simplifying yields

$$\mathbb{P}\big[\widetilde{w}^T\mathbf{Q}^T\widetilde{w} \geq 2|\!|\!|\mathbf{Q}|\!|\!|_{\mathrm{F}}^2 + 2\delta|\!|\!|\mathbf{Q}|\!|\!|_{\mathrm{op}}\big] \leq 2\exp\big(-\frac{2\delta}{|\!|\!|\mathbf{Q}|\!|\!|_{\mathrm{op}}}\big).$$

Putting together the pieces, we find that

$$\begin{aligned}\mathbb{P}[|Z| \geq \delta] &\leq 2\exp\big(-\frac{\delta^2}{4\,|\!|\!|\mathbf{Q}|\!|\!|_{\mathrm{F}}^2 + 4\delta|\!|\!|\mathbf{Q}|\!|\!|_{\mathrm{op}}}\big) + 2\exp\big(-\frac{2\delta}{|\!|\!|\mathbf{Q}|\!|\!|_{\mathrm{op}}}\big)\\ &\leq 4\exp\big(-\frac{\delta^2}{4\,|\!|\!|\mathbf{Q}|\!|\!|_{\mathrm{F}}^2 + 4\delta|\!|\!|\mathbf{Q}|\!|\!|_{\mathrm{op}}}\big).\end{aligned}$$

We have thus shown that the Gaussian chaos variable satisfies a sub-exponential tail bound. ♣

Example 2.16 (Singular values of Gaussian random matrices)**.** For integers $n > d$, let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be a random matrix with i.i.d. $\mathcal{N}(0,1)$ entries, and let

$$\gamma_1(\mathbf{X}) \;\geq\; \gamma_2(\mathbf{X}) \;\geq\; \ldots \;\geq\; \gamma_d(\mathbf{X}) \;\geq 0$$

denote its ordered singular values. By Wielandt's theorem (see Exercise 8.3), given another matrix $\mathbf{Y} \in \mathbb{R}^{n\times d}$, we have

$$\max_{k=1,\ldots,d} |\gamma_k(\mathbf{X}) - \gamma_k(\mathbf{Y})| \leq |\!|\!|\mathbf{X}-\mathbf{Y}|\!|\!|_{\mathrm{op}} \;\leq\; |\!|\!|\mathbf{X}-\mathbf{Y}|\!|\!|_{\mathrm{F}}. \tag{2.49}$$

where $|\!|\!|\cdot|\!|\!|_{\mathrm{F}}$ denotes the Frobenius norm on matrices—that is, the Euclidean norm on the matrix viewed as a vector in $\mathbb{R}^{nd}$. The inequality (2.49) shows that each singular value $\gamma_k(\mathbf{X})$ is a 1-Lipschitz function of the random matrix, so that Theorem 2.4 implies that for each $k = 1,\ldots,d$, we have

$$\mathbb{P}\big[\big|\gamma_k(\mathbf{X}) - \mathbb{E}[\gamma_k(\mathbf{X})]\big| \geq \sqrt{n}\delta\big] \leq 2e^{-\frac{n\delta^2}{2}} \quad \text{for all } \delta \geq 0. \tag{2.50}$$

Consequently, even though our techniques are not yet powerful enough to characterize the expected value of these random singular values, we are guaranteed that the expectations are representative of the typical behavior. See Chapter 6 for a more detailed discussion of the singular values of random matrices. ♣

Appendix A: Equivalent versions of sub-Gaussian variables

We establish the equivalence by proving the circle of implications (I) $\Rightarrow$ (II) $\Rightarrow$ (III) $\Rightarrow$ (I), followed by the equivalence (I) $\iff$ (IV).

Implication (I) $\Rightarrow$ (II)*:* If X is zero-mean and sub-Gaussian with parameter σ, then we claim that for $Z \sim \mathcal{N}(0, 2\sigma^2)$,

$$\frac{\mathbb{P}[X \geq t]}{\mathbb{P}[Z \geq t]} \leq \sqrt{8}\, e \qquad \text{for all } t \geq 0,$$

showing that X is majorized by Z with constant $c = \sqrt{8}\, e$. On one hand, by the sub-Gaussianity of X, we have $\mathbb{P}[X \geq t] \leq \exp(-\frac{t^2}{2\sigma^2})$ for all $t \geq 0$. On the other hand, by the Mill's ratio for Gaussian tails, if $Z \sim \mathcal{N}(0, 2\sigma^2)$, then we have

$$\mathbb{P}[Z \geq t] \geq \; \big(\frac{\sqrt{2}\sigma}{t} - \frac{(\sqrt{2}\sigma)^3}{t^3}\big)\, e^{-\frac{t^2}{4\sigma^2}} \qquad \text{for all } t > 0. \tag{2.51}$$

(See Exercise 2.2 for a derivation of this inequality.) We split the remainder of our analysis into two cases.

Case 1: First, suppose that $t \in [0, 2\sigma]$. Since the function $\Phi(t) = \mathbb{P}[Z \geq t]$ is decreasing, for all t in this interval,

$$\mathbb{P}[Z \geq t] \;\geq\; \mathbb{P}[Z \geq 2\sigma] \;\geq\; \big(\frac{1}{\sqrt{2}} - \frac{1}{2\sqrt{2}}\big)e^{-1} \;=\; \frac{1}{\sqrt{8}\, e}$$

Since $\mathbb{P}[X \geq t] \leq 1$, we conclude that $\frac{\mathbb{P}[X \geq t]}{\mathbb{P}[Z \geq t]} \leq \sqrt{8}\, e$ for all $t \in [0, 2\sigma]$.

Case 2: Otherwise, we may assume that $t > 2\sigma$. In this case, by combining the Mill's ratio (2.51) and the sub-Gaussian tail bound and making the substitution $s = t/\sigma$, we find that

$$\begin{aligned}
\sup_{t > 2\sigma} \frac{\mathbb{P}[X \geq t]}{\mathbb{P}[Z \geq t]} &\leq \sup_{s > 2} \frac{e^{-\frac{s^2}{4}}}{\big(\frac{\sqrt{2}}{s} - \frac{(\sqrt{2})^3}{s^3}\big)} \\
&\leq \sup_{s > 2} s^3\, e^{-\frac{s^2}{4}} \\
&\leq \sqrt{8} e,
\end{aligned}$$

where the last step follows from a numerical calculation.

Implication (II) $\Rightarrow$ (III)*:* Suppose that X is majorized by a zero-mean Gaussian with variance τ^2. Since X^{2k} is a non-negative random variable, we have

$$\mathbb{E}[X^{2k}] = \int_0^\infty \mathbb{P}[X^{2k} > s] ds \;=\; \int_0^\infty \mathbb{P}[|X| > s^{1/(2k)}]\, ds.$$

Under the majorization assumption, there is some constant $c \geq 1$ such that

$$\int_0^\infty \mathbb{P}[|X| > s^{1/(2k)}] ds \leq c \int_0^\infty \mathbb{P}[|Z| > s^{1/(2k)}] ds \;=\; c\,\mathbb{E}[Z^{2k}],$$

where $Z \sim \mathcal{N}(0, \tau^2)$. The polynomial moments of this centered Gaussian variable take the form,

$$\mathbb{E}[Z^{2k}] = \frac{(2k)!}{2^k k!} \tau^{2k}, \quad \text{for } k = 1, 2, \ldots, \tag{2.52}$$

whence

$$\mathbb{E}[X^{2k}] \leq c\,\mathbb{E}[Z^{2k}] = c\,\frac{(2k)!}{2^k k!} \tau^{2k} \;\leq\; \frac{(2k)!}{2^k k!} (c\,\tau)^{2k}, \qquad \text{for all } k = 1, 2, \ldots,$$

1 which establishes the moment bound (2.12) with $\theta = c\,\tau$.

2

Implication (III) $\Rightarrow$ (I)*:* For each $\lambda \in \mathbb{R}$, we have

$$\mathbb{E}[e^{\lambda X}] \leq 1 + \sum_{k=2}^\infty \frac{|\lambda|^k \mathbb{E}[|X|^k]}{k!}, \tag{2.53}$$

where we have used the fact $\mathbb{E}[X] = 0$ to elminate the term involving $k = 1$. If X is symmetric around zero, then all of its odd moments vanish, and by applying our assumption on $\theta(X)$, we obtain

$$\mathbb{E}[e^{\lambda X}] \leq 1 + \sum_{k=1}^\infty \frac{\lambda^{2k}}{(2k)!} \frac{(2k)!\,\theta^{2k}}{2^k k!} \;=\; e^{\frac{\lambda^2 \theta^2}{2}},$$

3 which shows that X is sub-Gaussian with parameter θ.

When X is not symmetric, we can bound the odd moments in terms of the even ones as

$$\mathbb{E}[|\lambda X|^{2k+1}] \stackrel{(i)}{\leq} \big(\mathbb{E}[|\lambda X|^{2k}]\, \mathbb{E}[|\lambda X|^{2k+2}]\big)^{1/2} \;\stackrel{(ii)}{\leq}\; \frac{1}{2}\big(\lambda^{2k} \mathbb{E}[X^{2k}] + \lambda^{2k+2} \mathbb{E}[X^{2k+2}]\big), \tag{2.54}$$

where step (i) follows from the Cauchy-Schwarz inequality; and step (ii) follows from the arithmetic-geometric mean inequality. Applying this bound to the power series expansion (2.53), we obtain

$$\begin{aligned}
\mathbb{E}[e^{\lambda X}] &\leq 1 + \big(\frac{1}{2} + \frac{1}{2 \cdot 3!}\big)\lambda^2 \mathbb{E}[X^2] + \sum_{k=2}^{\infty} \Big(\frac{1}{(2k)!} + \frac{1}{2}\Big[\frac{1}{(2k-1)!} + \frac{1}{(2k+1)!}\Big]\Big) \lambda^{2k} \mathbb{E}[X^{2k}] \\
&\leq \sum_{k=0}^{\infty} 2^k \frac{\lambda^{2k} \mathbb{E}[X^{2k}]}{(2k)!} \\
&\leq e^{\frac{(\sqrt{2}\lambda\theta)^2}{2}},
\end{aligned}$$

which establishes the claim.

(I) $\Rightarrow$ (IV) : This result is obvious for $s = 0$. For $s \in (0, 1)$, we begin with the sub-Gaussian inequality $\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\lambda^2 \sigma^2}{2}}$, and multiply both sides by $e^{-\frac{\lambda^2 \sigma^2}{2s}}$, thereby obtaining

$$\mathbb{E}\big[e^{\lambda X - \frac{\lambda^2 \sigma^2}{2s}}\big] \leq e^{\frac{\lambda^2 \sigma^2 \, (s-1)}{2s}}.$$

Since this inequality holds for all $\lambda \in \mathbb{R}$, we may integrate both sides over $\lambda \in \mathbb{R}$, using Fubini's theorem to justify exchanging the order of integration. On the right-hand side, we have

$$\int_{-\infty}^{\infty} \exp\Big(\frac{\lambda^2 \sigma^2 \, (s-1)}{2s}\Big) d\lambda = \frac{1}{\sigma} \sqrt{\frac{2\pi s}{1-s}}.$$

Turning to the left-hand side, for each fixed $x \in \mathbb{R}$, we have

$$\int_{-\infty}^{\infty} \exp\big(\lambda x - \frac{\lambda^2 \sigma^2}{2s}\big) d\lambda = \frac{\sqrt{2\pi s}}{\sigma} \; e^{\frac{s x^2}{2\sigma^2}}.$$

Taking expectations with respect to X, we conclude that

$$\mathbb{E}\big[e^{\frac{s X^2}{2\sigma^2}}\big] \leq \frac{\sigma}{\sqrt{2\pi s}} \, \frac{1}{\sigma} \sqrt{\frac{2\pi s}{1-s}} \; = \; \frac{1}{\sqrt{1-s}},$$

which establishes the claim.

(IV) $\Rightarrow$ (I)*:* From the bound $e^x \leq x + e^{9x^2/16}$, we have

$$\mathbb{E}[e^{\lambda X}] \leq \mathbb{E}[\lambda X] + \mathbb{E}\Big[e^{\frac{9\lambda^2 X^2}{16}}\Big] \; \stackrel{(i)}{\leq} \; e^{\frac{9\lambda^2 \sigma^2}{16}}, \tag{2.55}$$

where inequality (i) is valid for $|\lambda| \leq \frac{4}{3\sigma}$.

It remains to establish a similar upper bound for $|\lambda| > \frac{4}{3\sigma}$. Noting that for any $c > 0$, the functions $f(u) = \frac{u^2}{2c}$ and $f^*(v) = \frac{cv^2}{2}$ are conjugate duals, the Fenchel-Young inequality implies that

$$uv \leq \frac{u^2}{2c} + \frac{cv^2}{2}, \qquad \text{valid for all } u, v \in \mathbb{R} \text{ and } c > 0.$$

Applying this inequality with $u = \lambda$ and $v = X$ yields

$$\begin{aligned}
\mathbb{E}[e^{\lambda X}] &\leq \mathbb{E}\big[e^{\frac{\lambda^2\sigma^2}{2c} + \frac{cX^2}{2\sigma^2}}\big] \\
&= e^{\frac{\lambda^2\sigma^2}{2c}}\; \mathbb{E}\big[e^{\frac{cX^2}{2\sigma^2}}\big] \\
&\overset{(ii)}{\leq} e^{\frac{\lambda^2\sigma^2}{2c}}\; e^{c/2},
\end{aligned} \tag{2.56}$$

where the final inequality follows by Jensen's inequality, convexity of the exponential, and the fact that $\mathbb{E}[X^2] = \sigma^2$. Taking $c = 4/3$, then for all $|\lambda| \geq \frac{4}{3\sigma}$, we have

$$\frac{3\lambda^2\sigma^2}{8} \;=\; \frac{\lambda^2\sigma^2}{2c} = \frac{3\lambda^2\sigma^2}{8} \geq \frac{2}{3} = \frac{c}{2},$$

so that equation (2.56) implies that

$$\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\lambda^2\sigma^2}{c}} \;=\; e^{\frac{3\lambda^2\sigma^2}{4}} \qquad \text{for all } |\lambda| \geq \tfrac{4}{3\sigma}.$$

This inequality, combined with the bound (2.55), completes the proof.

Appendix B: Proof of Theorem 2.2

(II) $\Rightarrow$ (I)*:* The existence of the moment generating function for $|\lambda| < c_0$ implies that $\mathbb{E}[e^{\lambda X}] = 1 + \frac{\lambda^2 \mathbb{E}[X^2]}{2} + o(\lambda^2)$ as $\lambda \to 0$. Moreover, an ordinary Taylor series expansion implies that $e^{\frac{\sigma^2\lambda^2}{2}} = 1 + \frac{\sigma^2\lambda^2}{2} + o(\lambda^2)$ as $\lambda \to 0$. Therefore, as long as $\sigma^2 > \mathbb{E}[X^2]$, there exists some $b \geq 0$ such that $\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\sigma^2\lambda^2}{2}}$ for all $|\lambda| \leq \frac{1}{b}$.

(I) $\Rightarrow$ (II)*:* This implication is immediate.

(III) $\Rightarrow$ (II)*:* For an exponent $a > 0$ and truncation level $T > 0$ to be chosen, we have

$$\begin{aligned}
\mathbb{E}\big[e^{a|X|}\,\mathbb{I}[|X| \leq T]\big] &\leq \int_0^T \mathbb{P}[e^{a|X|} \geq t]dt \\
&\leq 1 + \int_1^T \mathbb{P}[|X| \geq \frac{\log t}{a}]dt.
\end{aligned}$$

Applying the assumed tail bound, we obtain

$$\begin{aligned}\mathbb{E}\big[e^{a|X|}\,\mathbb{I}[|X| \le T]\big] &\le 1 + c_1 \int_1^T e^{-\frac{c_2 \log t}{a}} dt \\ &= 1 + c_1 \int_1^T t^{-c_2/a} dt.\end{aligned}$$

For any $a \in [0, \frac{c_2}{2}]$, we have $\mathbb{E}\big[e^{a|X|}\,\mathbb{I}[|X| \le T]\big] \le 1 + \frac{c_1}{2}\big(1 - \frac{1}{T}\big)$, showing that $\mathbb{E}[e^{a|X|}]$ is finite for all $a \in [0, \frac{c_2}{2}]$. Since both e^{aX} and e^{-aX} are upper bounded by $e^{|a|\,|X|}$, we conclude that $\mathbb{E}[e^{aX}]$ is finite for all $|a| \le \frac{c_2}{2}$.

(II) $\Rightarrow$ (III)*:* By the Chernoff bound with $\lambda = c_0/2$, we have

$$\mathbb{P}[X \ge t] \le \mathbb{E}[e^{\frac{c_0 X}{2}}]\, e^{-\frac{c_0 t}{2}}$$

Applying a similar argument to $-X$, we conclude that $\mathbb{P}[|X| \ge t] \le c_1 e^{-c_2 t}$ with $c_1 = \mathbb{E}[e^{c_0 X/2}] + \mathbb{E}[e^{-c_0 X/2}]$ and $c_2 = c_0/2$.

(II) $\Longleftrightarrow$ (IV)*:* Since the moment generating function exists in an open interval around zero, we can consider the power series expansion

$$\mathbb{E}[e^{\lambda X}] = 1 + \sum_{k=2}^{\infty} \frac{\lambda^k \mathbb{E}[X^k]}{k!} \quad \text{for all } |\lambda| < a. \tag{2.57}$$

By definition, the quantity $\gamma(X)$ is the radius of convergence of this power series, from which the equivalence between (II) and (IV) follows.

■ 2.4 Bibliographic details and background

Further background and details on tail bounds can be found in various books (e.g., [BK00, Pet95, BLM13, SS91]). Classic papers on tail bounds include those of Bernstein [Ber37], Chernoff [Che52], Bahadur and Ranga [BR60], Bennett [Ben62], Hoeffding [Hoe63] and Azuma [Azu67]. The idea of using the cumulant function to bound the tails of a random variable was first introduced by Bernstein [Ber37], and further developed by Chernoff [Che52], whose name is now frequently associated with the method. The book [SS91] provides a number of more refined results that can be established using cumulant-based techniques. The original work of Hoeffding [Hoe63] gives results both for sums of independent random variables, assumed to be bounded from above, as well as certain types of dependent random variables, including U-statistics. The work of Azuma [Azu67] applies to general martingales that are sub-Gaussian in a conditional sense, as in Theorem 2.3.

The book by Buldygin and Kozachenko [BK00] provides a range of results on sub-Gaussian and sub-exponential variates. In particular, Theorems 2.1 and 2.2 are based on results from this book. The Orlicz norms, discussed briefly in Exercises 2.18 and 2.19, provide an elegant generalization of the sub-exponential and sub-Gaussian families. See Section 5.7 and the books [BK00, LT91] for further background on Orlicz norms.

The Johnson-Lindenstrauss lemma, discussed in Example 2.6, was originally proved in the paper [JL84] as an intermediate step in a more general result about Lipschitz embeddings. The original proof of lemma was based on random matrices with orthonormal rows, as opposed to the standard Gaussian random matrix used here. The use of random projection for dimension reduction and algorithmic speed-ups has a wide range of applications; see the book by Vempala [Vem04] for an overview.

Tail bounds for U-statistics, as sketched out in Example 2.10, were derived by Hoeffding [Hoe63]. The book by de la Peña and Giné [dLPG99] provides more advanced results, including extensions to uniform laws for U-processes and decoupling results. The bounded differences inequality (Corollary 2.2) and extensions thereof have many applications in the study of randomized algorithms as well as random graphs and other combinatorial objects. A number of such applications can be found in the survey by McDiarmid [McD89], and the book by Boucheron et al. [BLM13].

Milman and Schechtman [MS86] provide the short proof of Gaussian concentration for Lipschitz functions, on which Theorem 2.4 is based. Ledoux [Led01] provides an example of a Lipschitz function of i.i.d sequence of Rademacher variables, i.e., taking values $\{-1, +1\}$ equiprobably) for which sub-Gaussian concentration fails to hold (c.f. p. 128). However, sub-Gaussian concentration does hold for Lipschitz functions of bounded random variables with an additional convexity condition; see Section 3.3.5 for further details.

The kernel density estimation problem from Exercise 2.15 is a particular form of non-parametric estimation; we return to such problems in Chapters 13 and 14. Although we have focused exclusively on tail bounds for real-valued random variables, there are many generalizations to random variables taking values in Hilbert and other function spaces, as considered in Exercise 2.16. The books [LT91, Yur95] contain further background on such results. We also return to consider some versions of these bounds in Chapter 14. The Hanson-Wright inequality discussed in Exercise 2.17 was proved in the papers [HW71, Wri73]; see the papers [HKZ12, RV13] for more modern treatments. The moment-based tail bound from Exercise 2.20 relies on a classical inequality due to Rosenthal [Ros70]. Exercise 2.21 outlines the proof of the rate-distortion theorem for the Bernoulli source. It is a particular instance of more general information-theoretic results that are proved using probabilistic techniques: see the book by Cover and Thomas [CT91] for further reading. The Ising model (11.3) discussed in Exercise 2.22 has a lengthy history dating back to Ising [Isi25]. The book by Talagrand [Tal03] contains a wealth of information on spin glass models and their mathematical properties.

■ 2.5 Exercises

Exercise 2.1 (Tightness of inequalities)**.** The Markov and Chebyshev inequalities cannot be improved in general.

(a) Provide a random variable $X \geq 0$ for which Markov's inequality (2.1) is met with equality.

(b) Provide a random variable Y for which Chebyshev's inequality (2.2) is met with equality.

Exercise 2.2. Let $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ be the density function of a standard normal $Z \sim \mathcal{N}(0,1)$ variate.

(a) Show that $\phi'(z) + z\phi(z) = 0$.

(b) Use part (a) to show that

$$\phi(z)\left(\frac{1}{z} - \frac{1}{z^3}\right) \leq \mathbb{P}[Z \geq z] \leq \phi(z)\left(\frac{1}{z} - \frac{1}{z^3} + \frac{3}{z^5}\right) \quad \text{for all } z > 0. \tag{2.58}$$

Exercise 2.3. Suppose that $X \geq 0$, and that the moment generating function of X exists in an interval around zero. Given some $\delta > 0$ and integer $k = 1, 2, \ldots$, show that

$$\inf_{k=0,1,2,\ldots} \frac{\mathbb{E}[|X|^k]}{\delta^k} \leq \inf_{\lambda > 0} \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda \delta}}. \tag{2.59}$$

Consequently, an optimized bound based on polynomial moments is always at least as good as the Chernoff upper bound.

Exercise 2.4. Consider a random variable X with mean $\mu = \mathbb{E}[X]$, and such that $a \leq X \leq b$ almost surely.

(a) Defining the function $\psi(\lambda) = \log \mathbb{E}[e^{\lambda X}]$, show that $\psi(0) = 0$ and $\psi'(0) = \mu$.

(b) Show that $\psi''(\lambda) = \mathbb{E}_\lambda[X^2] - (\mathbb{E}_\lambda[X])^2$, where we define $\mathbb{E}_\lambda[f(X)] := \frac{\mathbb{E}[f(X)e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]}$. Use this fact to obtain an upper bound on $\sup_{\lambda \in \mathbb{R}} |\psi''(\lambda)|$.

(c) Use parts (a) and (b) to establish that X is sub-Gaussian with parameter at most $\sigma = \frac{b-a}{2}$.

Exercise 2.5 (Sub-Gaussian bounds and means/variances)**.** Consider a random variable X such that

$$\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\lambda^2\sigma^2}{2} + \lambda\mu} \tag{2.60}$$

for all $\lambda \in \mathbb{R}$.

(a) Show that $\mathbb{E}[X] = \mu$.

(b) Show that $\operatorname{var}(X) \leq \sigma^2$.

(c) Suppose that the smallest possible σ satisfying the inequality (2.60) is chosen. Is it then true that $\operatorname{var}(X) = \sigma^2$? Prove or disprove.

Exercise 2.6 (Lower bounds on squared sub-Gaussians)**.** Letting $\{X_i\}_{i=1}^n$ be an i.i.d. sequence of zero-mean sub-Gaussian variables with parameter σ, consider the normalized sum $Z_n := \frac{1}{n}\sum_{i=1}^n X_i^2$. Prove that

$$\mathbb{P}\big[Z_n - \mathbb{E}[Z_n] \leq \sigma^2\delta\big] \leq e^{-n\delta^2/16} \qquad \text{for all } \delta \geq 0.$$

This result shows that the lower tail of a sum of squared sub-Gaussian variables behaves in a sub-Gaussian way.

Exercise 2.7 (Bennett's inequality)**.** (a) Consider a zero-mean random variable such that $|X_i| \leq b$. Prove that

$$\log \mathbb{E}[e^{\lambda X_i}] \leq \sigma_i^2 \lambda^2 \, \Big\{\frac{e^{\lambda b} - 1 - \lambda b}{(\lambda b)^2}\Big\} \quad \text{for all } \lambda \in \mathbb{R},$$

where $\sigma_i^2 = \operatorname{var}(X_i)$.

(b) Given independent random variables $X_1, \ldots, X_n$ satisfying the condition of part (a), let $\sigma^2 := \frac{1}{n}\sum_{i=1}^n \sigma_i^2$ be the average variance. Prove *Bennett's inequality*

$$\mathbb{P}\big[\sum_{i=1}^n X_i \geq n\delta\big] \leq \exp\Big\{-\frac{n\sigma^2}{b^2} h\big(\frac{b\delta}{\sigma^2}\big)\Big\}, \tag{2.61}$$

where $h(t) := (1+t)\log(1+t) - t$ for $t \geq 0$.

(c) Show that Bennett's inequality is at least as good as Bernstein's inequality.

Exercise 2.8 (Bernstein and expectations)**.** Consider a non-negative random variable that satisfies a concentration inequality of the form

$$\mathbb{P}[Z \geq t] \leq C\, e^{-\frac{t^2}{2(\nu^2 + Bt)}} \tag{2.62}$$

for positive constants (ν, b) and $C \geq 1$.

(a) Show that $\mathbb{E}[Z] \leq 2\nu(\sqrt{\pi} + \sqrt{\log C}) + 4B(1 + \log C)$.

(b) Let $\{X_i\}_{i=1}^n$ be an i.i.d. sequence of zero-mean variables satisfying the Bernstein condition (2.16). Use part (a) to show that

$$\mathbb{E}\big[|\frac{1}{n}\sum_{i=1}^n X_i|\big] \leq \frac{2\sigma}{\sqrt{n}}\big(\sqrt{\pi} + \sqrt{\log 2}\big) + \frac{4b}{n}\big(1 + \log 2\big),$$

Exercise 2.9 (Sharp upper bounds on binomial tails)**.** Let $\{X_i\}_{i=1}^n$ be an i.i.d. sequence of Bernoulli variables with parameter $\alpha \in (0, 1/2]$, and consider the binomial random variable $Z_n = \sum_{i=1}^n X_i$. The goal of this exercise is to prove, for any $\delta \in (0, \alpha)$, a sharp upper bound on the tail probability $\mathbb{P}[Z_n \leq \delta n]$.

(a) Show that $\mathbb{P}[Z_n \leq \delta n] \leq e^{-nD(\delta \,\|\, \alpha)}$, where the quantity

$$D(\delta \,\|\, \alpha) := \delta \log \frac{\delta}{\alpha} + (1-\delta) \log \frac{(1-\delta)}{(1-\alpha)} \tag{2.63}$$

is the Kullback-Leibler divergence between the Bernoulli distributions with parameters δ and α respectively.

(b) Show that the bound from part (a) is strictly better than the Hoeffding bound for all $\delta \in (0, \alpha)$.

Exercise 2.10 (Lower bounds on binomial tails)**.** Let $\{X_i\}_{i=1}^n$ be a sequence of i.i.d. Bernoulli variables with parameter $\alpha \in (0, 1/2]$, and consider the binomial random variable $Z_n = \sum_{i=1}^n X_i$. In this exercise, we establish a *lower bound* on the probability $\mathbb{P}[Z_n \leq \delta n]$ for some fixed $\delta \in (0, \alpha)$, thereby establishing that the upper bound from Exercise 2.9 is tight up to a polynomial pre-factor. Throughout the analysis, we define $m = \lfloor n\delta \rfloor$, the largest integer less than or equal to $n\delta$, and set $\widetilde{\delta} = \frac{m}{n}$.

(a) Prove $\frac{1}{n} \log \mathbb{P}[Z_n \leq \delta n] \geq \frac{1}{n} \log \binom{n}{m} + \widetilde{\delta} \log \alpha + (1 - \widetilde{\delta}) \log(1 - \alpha)$.

(b) Show that

$$\frac{1}{n} \log \binom{n}{m} \geq \phi(\widetilde{\delta}) - \frac{\log(n+1)}{n} \tag{2.64}$$

where $\phi(\widetilde{\delta}) = -\widetilde{\delta} \log(\widetilde{\delta}) - (1-\widetilde{\delta}) \log(1-\widetilde{\delta})$ is the binary entropy. (*Hint:* Let Y be binomial random variable with parameters $(n, \widetilde{\delta})$ and show that $\mathbb{P}[Y = \ell]$ is maximized when $\ell = m = \widetilde{\delta} n$.)

(c) Show that

$$\mathbb{P}[Z_n \le \delta n] \ge \frac{1}{n+1}\, e^{-nD(\delta \,\|\, \alpha)}, \tag{2.65}$$

where the Kullback-Leibler divergence $D(\delta \,\|\, \alpha)$ was previously defined (2.63).

Exercise 2.11 (Gaussian maxima)**.** Let $\{X_i\}_{i=1}^n$ be an i.i.d. sequence of $\mathcal{N}(0, \sigma^2)$ variables, and consider the random variable $Z_n := \max_{i=1,\ldots,n} |X_i|$.

(a) Prove that

$$\mathbb{E}[Z_n] \le \sqrt{2\sigma^2 \log n} + \frac{4\sigma}{\sqrt{2 \log n}} \quad \text{for all } n \ge 2.$$

(*Hint:* You may use the tail bound $\mathbb{P}[U \ge \delta] \le \sqrt{\frac{2}{\pi}}\, \frac{1}{\delta} e^{-\delta^2/2}$, valid for any standard normal variate.)

(b) Prove that

$$\mathbb{E}[Z_n] \ge (1 - 1/e)\, \sqrt{2\sigma^2 \log n} \quad \text{for all } n \ge 5.$$

(c) Prove that $\frac{\mathbb{E}[Z_n]}{\sqrt{2\sigma^2 \log n}} \to 1$ as $n \to +\infty$.

Exercise 2.12 (Sharp upper bounds for sub-Gaussian maxima)**.** Let $\{X_i\}_{i=1}^n$ be a sequence of zero-mean random variables, each sub-Gaussian with parameter σ. (No independence assumptions are needed.) Prove that

$$\mathbb{E}\big[\max_{i=1,\ldots,n} X_i\big] \le \sqrt{2\sigma^2 \log n} \quad \text{for all } n \ge 1. \tag{2.66}$$

(*Hint:* The exponential is a convex function.)

Exercise 2.13 (Operations on sub-Gaussian variables)**.** Suppose that X_1 and X_2 are zero-mean and sub-Gaussian with parameters σ_1 and σ_2 respectively.

(a) If X_1 and X_2 are independent, show that $X_1 + X_2$ is sub-Gaussian with parameter $\sqrt{\sigma_1^2 + \sigma_2^2}$.

(b) Show that in general (without assuming independence), the random variable $X_1 + X_2$ is sub-Gaussian with parameter at most $2\sqrt{\sigma_1^2 + \sigma_2^2}$.

(c) If X_1 and X_2 are independent, show that $X_1 X_2$ is sub-exponential with parameters $(\nu, b) = (\sqrt{2}\sigma_1\sigma_2,\ \frac{1}{\sqrt{2}\sigma_1\sigma_2})$.

Exercise 2.14 (Concentration around medians and means)**.** Given a scalar random variable X, suppose that there are positive constants c_1, c_2 such that

$$\mathbb{P}\big[|X - \mathbb{E}[X]| \geq t\big] \leq c_1 e^{-c_2 t^2} \quad \text{for all } t \geq 0. \tag{2.67}$$

(a) Prove that $\text{var}(X) \leq \frac{c_1}{c_2}$.

(b) A median m_X is any number such that $\mathbb{P}[X \geq m_X] \geq 1/2$ and $\mathbb{P}[X \leq m_X] \geq 1/2$. Show by example that the median need not be unique.

(c) Show that whenever the mean concentration bound (2.67) holds, then for any median m_X,

$$\mathbb{P}\big[|X - m_X| \geq t\big] \leq c_3 e^{-c_4 t^2} \qquad \text{for all } t \geq 0, \tag{2.68}$$

where $c_3 := 4c_1$ and $c_4 := \frac{c_2}{8}$.

(d) Conversely, show that whenever the median concentration bound (2.68) holds, then mean concentration (2.67) holds with $c_1 = 2c_3$ and $c_2 = \frac{c_4}{4}$.

Exercise 2.15 (Concentration and kernel density estimation)**.** Let $X_1, X_2, \ldots, X_n$ be i.i.d. samples of random variable with density f on the real line. A standard estimate of f is the *kernel density estimate*

$$f_n(x) := \frac{1}{nh} \sum_{i=1}^{n} K\big(\frac{x - X_i}{h}\big),$$

where $K : \mathbb{R} \to [0, \infty)$ is a kernel function satisfying $\int_{-\infty}^{\infty} K(t)dt = 1$, and $h > 0$ is a bandwidth parameter. Suppose that we assess the quality of $\widehat{f}_n$ using the L^1 norm $\|\widehat{f}_n - f\|_1 := \int_{-\infty}^{\infty} |\widehat{f}_n(t) - f(t)|dt$. Prove that

$$\mathbb{P}\big[\|\widehat{f}_n - f\|_1 \geq \mathbb{E}[\|\widehat{f}_n - f\|_1] + \delta\big] \leq e^{-\frac{n\delta^2}{8}}.$$

Exercise 2.16 (Deviation inequalities in a Hilbert space)**.** Let $\{X_i\}_{i=1}^n$ be a sequence of independent random variables taking values in a Hilbert space $\mathbb{H}$, and suppose that $\|X_i\|_{\mathbb{H}} \leq b_i$ almost surely. Consider the real-valued random variable $S_n = \|\sum_{i=1}^n X_i\|_{\mathbb{H}}$.

(a) Show that for all $\delta \geq 0$,

$$\mathbb{P}\big[|S_n - \mathbb{E}[S_n]| \geq n\delta\big] \leq 2e^{-\frac{n\delta^2}{8b^2}}, \quad \text{where } b^2 = \frac{1}{n}\sum_{i=1}^n b_i^2.$$

(b) Show that $\mathbb{P}\big[\frac{S_n}{n} \geq a + \delta\big] \leq e^{-\frac{n\delta^2}{8b^2}}$, where $a := \sqrt{\frac{1}{n^2}\sum_{i=1}^n \mathbb{E}[\|X_i\|_{\mathbb{H}}^2]}$.

Exercise 2.17 (Hanson-Wright inequality)**.** Given a positive semidefinite matrix $\mathbf{Q} \in \mathbb{R}^{n\times n}$, consider the random variable $Z = \sum_{i=1}^n \sum_{j=1}^n \mathbf{Q}_{ij} X_i X_j$. The *Hanson-Wright inequality* guarantees that whenever the random variables $\{X_i\}_{i=1}^n$ are i.i.d. variables with mean zero, variance 1 and σ-sub-Gaussian, then there are universal constants (c_1, c_2) such that

$$\mathbb{P}\big[Z \geq \operatorname{trace}(\mathbf{Q}) + \sigma t\big] \leq 2e^{-\min\big\{\frac{c_1 t}{\|\mathbf{Q}\|_{\text{op}}}, \frac{c_2 t^2}{\|\mathbf{Q}\|_{\text{F}}^2}\big\}}, \tag{2.69}$$

where $\|\mathbf{Q}\|_{\text{op}}$ and $\|\mathbf{Q}\|_{\text{F}}$ denote the operator and Frobenius norms, respectively. Prove this inequality in the special case $X_i \sim N(0,1)$. (*Hint:* The rotation-invariance of the Gaussian distribution and sub-exponential nature of χ^2-variates could be useful.)

Exercise 2.18 (Orlicz norms)**.** Let $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ be a strictly increasing convex function that satisfies $\psi(0) = 0$. The ψ-*Orlicz* norm of a random variable X is defined as

$$\|X\|_\psi := \inf\big\{t > 0 \mid \mathbb{E}\big[\psi\big(t^{-1}|X|\big)\big] \leq 1\big\}, \tag{2.70}$$

where $\|X\|_\psi$ is infinite if there is no finite t for which the expectation $\mathbb{E}[\psi(t^{-1}|X|)]$ exists. For the functions $u \mapsto u^q$ for some $q \in [1, \infty]$, then the Orlicz norm is simply the usual ℓ_q-norm $\|X\|_q = \big(\mathbb{E}[|X|^q]\big)^{1/q}$. In this exercise, we consider the Orlicz norms $\|\cdot\|_{\psi_q}$ defined by the convex functions $\psi_q(u) = \exp(u^q) - 1$, for $q \geq 1$.

(a) If $\|X\|_{\psi_q} < +\infty$, show that there exist positive constants c_1, c_2 such that

$$\mathbb{P}[|X| > t] \leq c_1 \exp(-c_2 t^q) \qquad \text{for all } t > 0. \tag{2.71}$$

(In particular, you should be able to show that this bound holds with $c_1 = 2$ and $c_2 = \|X\|_{\psi_q}^{-q}$.)

(b) Suppose that a random variable Z satisfies the tail bound (2.71). Show that $\|X\|_{\psi_q}$ is finite.

Exercise 2.19 (Maxima of Orlicz variables)**.** Recall the definition of Orlicz norm from Exercise 2.18. Let $\{X_i\}_{i=1}^n$ be an i.i.d. sequence of zero-mean random variables with finite Orlicz norm $\sigma = \|X_i\|_\psi$. Show that

$$\mathbb{E}[\max_{i=1,\ldots,n} |X_i|] \leq \sigma\psi^{-1}(n).$$

Exercise 2.20 (Tail bounds under moment conditions)**.** Suppose that $\{X_i\}_{i=1}^n$ are zero-mean and independent random variables such that, for some fixed integer $m \geq 1$, they satisfy the moment bound $\|X_i\|_{2m} := (\mathbb{E}[X_i^{2m}])^{\frac{1}{2m}} \leq C_m$

$$\mathbb{P}\big[|\frac{1}{n}\sum_{i=1}^n X_i| \geq \delta\big] \leq B_m\big(\frac{1}{\sqrt{n}\delta}\big)^{2m} \quad \text{for all } \delta > 0,$$

where B_m is a universal constant depending only on C_m and m.
Hint: You may find the following form of Rosenthal's inequality to be useful: under the stated conditions, there is a universal constant R_m such that

$$\mathbb{E}\big[\big(\sum_{i=1}^n X_i\big)^{2m}\big] \leq R_m\Big\{\sum_{i=1}^n \mathbb{E}[X_i^{2m}] + \big(\sum_{i=1}^n \mathbb{E}[X_i^2]\big)^m\Big\}.$$

Exercise 2.21 (Concentration and data compression)**.** Let $X = (X_1, X_2, \ldots, X_n)$ be a vector of i.i.d. Bernoulli variables with parameter 1/2. The goal of lossy data compression is to represent X using a codebook of binary vectors, say $\{z^1, \ldots, z^N\}$, such that the *rescaled Hamming distortion*

$$\delta := \mathbb{E}\big[\min_{j=1,\ldots,N} \rho_H(X, z^j)\big] \;=\; \mathbb{E}\big[\min_{j=1,\ldots,N} \frac{1}{n}\sum_{i=1}^n \mathbb{I}[X_i \neq z_i^j]\big]$$

is as small as possible. Of course, one can always achieve zero distortion using a codebook with $N = 2^n$ codewords, so the goal is to use $N = 2^{Rn}$ codewords for some rate $R < 1$. In this exercise, we use tail bounds to study the trade-off between the rate R and the distortion δ.

(a) Suppose that the rate R is upper bounded as

$$R \;< D_2(\delta \,\|\, 1/2) = \delta \log_2 \frac{\delta}{1/2} + (1-\delta)\log_2 \frac{1-\delta}{1/2}.$$

Show that for any codebook $\{z^1, \ldots, z^N\}$, the probability of achieving distortion δ goes to zero as n goes to infinity. (*Hint:* Let V^j be an 0-1-valued indicator variable for the event $\rho_H(X, z^j) \leq \delta$, and define $V = \sum_{j=1}^N V^j$. The tail bounds from Exercise 2.9 could be useful in bounding the probability $\mathbb{P}[V \geq 1]$.)

(b) We now show that if $\Delta R := R - D_2(\delta \,\|\, 1/2) > 0$, then there exists a codebook that achieves distortion δ. In order to do so, consider a random codebook $\{Z^1, \ldots, Z^N\}$, formed by generating each codeword Z^j independently, and with all i.i.d. Ber(1/2) entries. Let V^j be an indicator for the event $\rho_H(X, Z^j) \leq \delta$, and define $V = \sum_{j=1}^N V^j$.

(i) Show that $\mathbb{P}[V \geq 1] \geq \frac{(\mathbb{E}[V])^2}{\mathbb{E}[V^2]}$.

(ii) Use part (i) to show that $\mathbb{P}[V \geq 1] \rightarrow +\infty$ as $n \rightarrow +\infty$. (*Hint:* The tail bounds from Exercise 2.10 could be useful.)

Exercise 2.22 (Concentration for spin glasses)**.** For some positive integer $d \geq 2$, consider a collection $\{\theta_{jk}\}_{j \neq k}$ of weights, one for each distinct pair $j \neq k$ of indices in $\{1, 2, \ldots, d\}$. We can then define a probability distribution over the Boolean hypercube $\{-1, +1\}^d$ via the mass function

$$\mathbb{P}_\theta(x_1, \ldots, x_d) = \exp\Big\{\frac{1}{\sqrt{d}} \sum_{i \neq j} \theta_{jk} x_j x_k \; - \; F_d(\theta)\Big\}, \tag{2.72}$$

where the function $F_d : \mathbb{R}^{\binom{d}{2}} \rightarrow \mathbb{R}$, known as the *free energy*, is given by

$$F_d(\theta) := \log\Bigg(\sum_{x \in \{-1,+1\}^d} \exp\Big\{\frac{1}{\sqrt{d}} \sum_{j \neq k} \theta_{jk} x_j x_k\Big\}\Bigg) \tag{2.73}$$

serves to normalize the distribution. The probability distribution (2.72) was originally used to describe the behavior of magnets in statistical physics, in which context it is known as the *Ising model*. Suppose that the weights are chosen as i.i.d. random variables, so that equation (2.72) now describes a random family of probability distributions. This family is known as the Sherrington-Kirkpatrick model in statistical physics.

(a) Show that F_d is a convex function.

(b) For any two vectors $\theta, \theta' \in \mathbb{R}^{\binom{d}{2}}$, show that $\|F_d(\theta) - F_d(\theta')\|_2 \leq \sqrt{d}\, \|\theta - \theta'\|_2$.

(c) Suppose that the weights are chosen in an i.i.d. manner as $\theta_{jk} \sim \mathcal{N}(0, \beta^2)$ for each $j \neq k$. Use the previous parts and Jensen's inequality to show that

$$\mathbb{P}\left[\frac{F_d(\theta)}{d} \geq \log 2 + \frac{\beta^2}{4} + t\right] \leq 2e^{-\beta d t^2/2} \quad \text{for all } t > 0. \tag{2.74}$$

Remark: Interestingly, it is known that for any $\beta \in [0, 1)$, this tail bound captures the behavior of $F_d(\theta)/d$ accurately, in that $\frac{F_d(\theta)}{d} \xrightarrow{a.s.} \log 2 + \beta^2/4$. In contrast, for $\beta \geq 1$, the behavior of this spin glass model is much more subtle; we refer the reader to the bibliographic section for additional reading.