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. This chapter serves as an entry point 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

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

This is a simple instance of an upper tail bound. For a random variable X that also has a finite variance, we have *Chebyshev's inequality*:

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

This is a simple form of concentration inequality, guaranteeing that X is close to its mean $\mu = \mathbb{E}[X]$ whenever its variance is small. Observe that Chebyshev's inequality follows by applying Markov's inequality to the non-negative random variable $Y = (X - \mu)^2$. Both Markov's and Chebyshev's inequalities 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

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

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

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 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

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

As we explore in Exercise 2.3, the moment bound (2.3) with an 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 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

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

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

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

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*

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

In fact, this bound is sharp up to polynomial-factor corrections, as shown by our exploration of the Mills ratio in Exercise 2.2. ♣

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

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

Definition 2.2 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*

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

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

Example 2.3 (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

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

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.

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

Example 2.4 (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

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

where the inequality follows from 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

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

where step (i) follows from the result of Example 2.3, applied conditionally with (X, X') held fixed. Since $|X - X'| \leq b - a$, we are guaranteed that

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

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 2.4 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 and X_2 are independent sub-Gaussian variables with parameters σ_1 and σ_2, then $X_1 + X_2$ is sub-Gaussian with parameter $\sqrt{\sigma_1^2 + \sigma_2^2}$. See Exercise 2.13 for verification of this fact, as well as some 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 2.5 (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*

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

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

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

sub-Gaussian with parameter $\sigma = \frac{b-a}{2}$, so that we obtain the bound

$$\mathbb{P}\left[\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right] \leq e^{-\frac{2t^2}{n(b-a)^2}}. \tag{2.11}$$

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

(I) *There is a constant $\sigma \geq 0$ such that*

$$\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\lambda^2\sigma^2}{2}} \qquad \text{for all } \lambda \in \mathbb{R}. \tag{2.12a}$$

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

$$\mathbb{P}[|X| \geq s] \leq c\,\mathbb{P}[|Z| \geq s] \qquad \text{for all } s \geq 0. \tag{2.12b}$$

(III) *There is a constant $\theta \geq 0$ such that*

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

(IV) *There is a constant $\sigma \geq 0$ such that*

$$\mathbb{E}[e^{\frac{\lambda X^2}{2\sigma^2}}] \leq \frac{1}{\sqrt{1-\lambda}} \qquad \text{for all } \lambda \in [0, 1). \tag{2.12d}$$

See Appendix A (Section 2.4) 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:

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

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

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{\nu^2\lambda^2}{2}} \qquad \text{for all } |\lambda| < \tfrac{1}{\alpha}. \tag{2.13}$$

It follows immediately from this definition that any sub-Gaussian variable is also sub-exponential—in particular, with $\nu = \sigma$ and $\alpha = 0$, 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 2.8 (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 $\lambda > \frac{1}{2}$, the moment generating function is infinite, which reveals 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 (2.13) is satisfied. Following some calculus, we find that

$$\frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \leq e^{2\lambda^2} = e^{4\lambda^2/2}, \qquad \text{for all } |\lambda| < \tfrac{1}{4}, \tag{2.14}$$

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

As with sub-Gaussianity, the control (2.13) 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 2.9 (Sub-exponential tail bound) *Suppose that X is sub-exponential with parameters (ν, α). Then*

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

As with the Hoeffding inequality, similar bounds can be derived for the left-sided event $\{X - \mu \leq -t\}$, as well as the two-sided event $\{|X - \mu| \geq t\}$, with an additional factor of 2 in the latter case.

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

Proof By recentering as needed, we may assume without loss of generality that $\mu = 0$. We follow the usual Chernoff-type approach: combining it with the definition (2.13) of a sub-exponential variable yields the upper bound

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

In order to complete the proof, it remains to compute, for each fixed $t \geq 0$, the quantity $g^*(t) := \inf_{\lambda\in[0,\alpha^{-1})} g(\lambda, t)$. Note that the unconstrained minimum of the function $g(\cdot, t)$ occurs at $\lambda^* = t/\nu^2$. If $0 \leq t < \frac{\nu^2}{\alpha}$, 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 $t \geq \frac{\nu^2}{\alpha}$. 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 $\lambda^\dagger = \alpha^{-1}$, and we have

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

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

As shown in Example 2.8, the sub-exponential property can be verified by explicitly computing or bounding the moment generating function. This direct calculation may be impracticable in many settings, so it is natural to seek alternative approaches. One such method is based on control of 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

$$|\mathbb{E}[(X-\mu)^k]| \leq \tfrac{1}{2}k!\sigma^2 b^{k-2} \qquad \text{for } k = 2, 3, 4, \ldots. \tag{2.15}$$

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 (2.15) holds. Even for bounded variables, our next result will show that the Bernstein condition can be used to obtain tail bounds that may be tighter than the Hoeffding bound. Moreover, Bernstein's condition is also satisfied by various unbounded variables, a property which lends 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

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

where the inequality (i) makes use of the Bernstein condition (2.15). For any $|\lambda| < 1/b$, we

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

can sum the geometric series so as to obtain

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq 1 + \frac{\lambda^2\sigma^2/2}{1-b|\lambda|} \stackrel{\text{(ii)}}{\leq} e^{\frac{\lambda^2\sigma^2/2}{1-b|\lambda|}}, \tag{2.16}$$

where inequality (ii) follows from the bound $1 + t \leq e^t$. 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 2.9 leads directly to tail bounds on a random variable satisfying the Bernstein condition (2.15). However, the resulting tail bound can be sharpened slightly, at least in terms of constant factors, by making direct use of the upper bound (2.16). We summarize in the following:

Proposition 2.10 (Bernstein-type bound) *For any random variable satisfying the Bernstein condition* (2.15)*, we have*

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{\lambda^2\sigma^2/2}{1-b|\lambda|}} \qquad \textit{for all } |\lambda| < \tfrac{1}{b}, \tag{2.17a}$$

and, moreover, the concentration inequality

$$\mathbb{P}[|X-\mu| \geq t] \leq 2e^{-\frac{t^2}{2(\sigma^2+bt)}} \qquad \textit{for all } t \geq 0. \tag{2.17b}$$

We proved inequality (2.17a) in the discussion preceding this proposition. Using this bound on the moment generating function, the tail bound (2.17b) 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 2.10 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 2.5). Alternatively, using the fact that any bounded variable satisfies the Bernstein condition (2.16), we can also apply Proposition 2.10, thereby obtaining the tail bound (2.17b), 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 (ν, α) transform in a simple

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

way. In particular, consider an independent sequence $\{X_k\}_{k=1}^n$ of random variables, such that X_k has mean μ_k, and is sub-exponential with parameters (ν_k, α_k). We compute the moment generating function

$$\mathbb{E}[e^{\lambda \sum_{k=1}^n (X_k - \mu_k)}] \stackrel{\text{(i)}}{=} \prod_{k=1}^n \mathbb{E}[e^{\lambda(X_k - \mu_k)}] \stackrel{\text{(ii)}}{\leq} \prod_{k=1}^n e^{\lambda^2 \nu_k^2/2},$$

valid for all $|\lambda| < \left(\max_{k=1,\ldots,n} \alpha_k\right)^{-1}$, where equality (i) follows from independence, and inequality (ii) follows since X_k is sub-exponential with parameters (ν_k, α_k). Thus, we conclude that the variable $\sum_{k=1}^n (X_k - \mu_k)$ is sub-exponential with the parameters (ν_*, α_*), where

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

Using the same argument as in Proposition 2.9, this observation leads directly to the upper tail bound

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n (X_k - \mu_k) \geq t\right] \leq \begin{cases} e^{-\frac{nt^2}{2(\nu_*^2/n)}} & \text{for } 0 \leq t \leq \frac{\nu_*^2}{n\alpha_*}, \\ e^{-\frac{nt}{2\alpha_*}} & \text{for } t > \frac{\nu_*^2}{n\alpha_*}, \end{cases} \tag{2.18}$$

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

Example 2.11 (χ^2-variables) A chi-squared (χ^2) 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 2.8, the variable Z_k^2 is sub-exponential with parameters $(2, 4)$. Consequently, since the variables $\{Z_k\}_{k=1}^n$ are independent, the χ^2-variate Y is sub-exponential with parameters $(\nu, \alpha) = (2\sqrt{n}, 4)$, and the preceding discussion yields the two-sided tail bound

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{k=1}^n Z_k^2 - 1\right| \geq t\right] \leq 2e^{-nt^2/8}, \qquad \text{for all } t \in (0,1). \tag{2.19}$$

♣

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 2.12 (Johnson–Lindenstrauss embedding) As one application of the concentration of χ^2-variables, consider the following problem. Suppose that we are given $N \geq 2$ distinct vectors $\{u^1, \ldots, u^N\}$, with each vector lying in $\mathbb{R}^d$. If the data dimension d is large, then it might be expensive to store and manipulate the data set. The idea of dimensionality reduction is to construct a mapping $F : \mathbb{R}^d \to \mathbb{R}^m$—with the projected dimension m substantially smaller than d—that preserves some "essential" features of the data set. What features should we try to preserve? There is not a unique answer to this question but, as one interesting example, we might consider preserving pairwise distances, or equivalently norms

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

and inner products. Many algorithms are based on such pairwise quantities, including linear regression, methods for principal components, the k-means algorithm for clustering, and nearest-neighbor algorithms for density estimation. With these motivations in mind, given some tolerance $\delta \in (0, 1)$, we might be interested in a mapping F with the guarantee that

$$(1 - \delta) \leq \frac{\|F(u^i) - F(u^j)\|_2^2}{\|u^i - u^j\|_2^2} \leq (1 + \delta) \qquad \text{for all pairs } u^i \neq u^j. \tag{2.20}$$

In words, the projected data set $\{F(u^1), \ldots, F(u^N)\}$ preserves all pairwise squared distances up to a multiplicative factor of δ. Of course, this is always possible if the projected dimension m is large enough, but the goal is to do it with relatively small m.

Constructing such a mapping that satisfies the condition (2.20) with high probability turns out to be straightforward as long as the projected dimension is lower bounded as $m \gtrsim \frac{1}{\delta^2} \log N$. Observe that the projected dimension is independent of the ambient dimension d, and scales only logarithmically with the number of data points N.

The construction is probabilistic: first form a random matrix $\mathbf{X} \in \mathbb{R}^{m \times d}$ filled with independent $\mathcal{N}(0, 1)$ entries, and use it to define a linear mapping $F : \mathbb{R}^d \to \mathbb{R}^m$ via $u \mapsto \mathbf{X}u/\sqrt{m}$. We now verify that F satisfies condition (2.20) with high probability. Let $x_i \in \mathbb{R}^d$ denote the ith row of $\mathbf{X}$, 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

$$Y := \frac{\|\mathbf{X}u\|_2^2}{\|u\|_2^2} = \sum_{i=1}^{m} \langle x_i, u/\|u\|_2 \rangle^2 ,$$

follows a χ^2 distribution with m degrees of freedom, using the independence of the rows. Therefore, applying the tail bound (2.19), we find that

$$\mathbb{P}\left[\left| \frac{\|\mathbf{X}u\|_2^2}{m \, \|u\|_2^2} - 1 \right| \geq \delta \right] \leq 2e^{-m\delta^2/8} \qquad \text{for all } \delta \in (0, 1).$$

Rearranging and recalling the definition of F yields the bound

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

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

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

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

In parallel to Theorem 2.13, there are a number of equivalent ways to characterize a subexponential random variable. The following theorem provides a summary:

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

Theorem 2.13 (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, α) such that*

$$\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\nu^2\lambda^2}{2}} \qquad \text{for all } |\lambda| < \tfrac{1}{\alpha}. \tag{2.21a}$$

(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*

$$\mathbb{P}[|X| \geq t] \leq c_1 e^{-c_2 t} \qquad \text{for all } t > 0. \tag{2.21b}$$

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

See Appendix B (Section 2.5) 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 yields bounds on both the upper and lower tails. As we have seen, one sufficient condition for Bernstein's condition to hold is a bound on the absolute value, 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 2.14 (One-sided Bernstein's inequality) *If $X \leq b$ almost surely, then*

$$\mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}] \leq \exp\left(\frac{\frac{\lambda^2}{2}\mathbb{E}[X^2]}{1 - \frac{b\lambda}{3}}\right) \qquad \text{for all } \lambda \in [0, 3/b). \tag{2.22a}$$

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

$$\mathbb{P}\left[\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \geq n\delta\right] \leq \exp\left(-\frac{n\delta^2}{2\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i^2] + \frac{b\delta}{3}\right)}\right). \tag{2.22b}$$

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 (2.22b) to the random variable $-X$. In the special case of independent non-negative random variables $Y_i \geq 0$, we find that

$$\mathbb{P}\left[\sum_{i=1}^n (Y_i - \mathbb{E}[Y_i]) \leq -n\delta\right] \leq \exp\left(-\frac{n\delta^2}{\frac{2}{n}\sum_{i=1}^n \mathbb{E}[Y_i^2]}\right). \tag{2.23}$$

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.

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

The proof of Proposition 2.14 is quite straightforward given our development thus far.

Proof Defining the function

$$h(u) := 2\frac{e^u - u - 1}{u^2} = 2\sum_{k=2}^{\infty} \frac{u^{k-2}}{k!},$$

we have the expansion

$$\mathbb{E}[e^{\lambda X}] = 1 + \lambda\mathbb{E}[X] + \tfrac{1}{2}\lambda^2\mathbb{E}[X^2 h(\lambda X)].$$

Observe that for all scalars $x < 0$, $x' \in [0, b]$ 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

$$\begin{aligned}\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] &\leq e^{-\lambda\mathbb{E}[X]}\{1 + \lambda\mathbb{E}[X] + \tfrac{1}{2}\lambda^2\mathbb{E}[X^2]h(\lambda b)\} \\ &\leq \exp\left\{\frac{\lambda^2\mathbb{E}[X^2]}{2}\, h(\lambda b)\right\}.\end{aligned}$$

Consequently, the bound (2.22a) 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

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

where the condition $\frac{\lambda b}{3} \in [0, 1)$ allows us to sum the geometric series.

In order to prove the upper tail bound (2.22b), we apply the Chernoff bound, exploiting independence to apply the moment generating function bound (2.22a) separately, and thereby find that

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

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.

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

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 on 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

$$Y_k = \mathbb{E}[f(X) \mid X_1, \ldots, X_k] \qquad \text{for } k = 1, \ldots, n-1, \tag{2.24}$$

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

$$f(X) - \mathbb{E}[f(X)] = Y_n - Y_0 = \sum_{k=1}^{n} \underbrace{(Y_k - Y_{k-1})}_{D_k},$$

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 2.15 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 $k \geq 1$,

$$\mathbb{E}[|Y_k|] < \infty \quad \text{and} \quad \mathbb{E}[Y_{k+1} \mid \mathcal{F}_k] = Y_k. \tag{2.25}$$

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 2.16 (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

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

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 $\{X_k\}_{k=1}^\infty$. ♣

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

Example 2.17 (Doob construction) Given a sequence of independent random variables $\{X_k\}_{k=1}^n$, 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, in terms of the shorthand $X_1^k = (X_1, X_2, \ldots, X_k)$, we have

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

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

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

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

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

Example 2.18 (Likelihood ratio) Let f and g be two mutually absolutely continuous densities, and let $\{X_k\}_{k=1}^\infty$ be a sequence of random variables drawn i.i.d. according to f. For each $k \geq 1$, let $Y_k := \prod_{\ell=1}^k \frac{g(X_\ell)}{f(X_\ell)}$ be the likelihood ratio based on the first k samples. Then the sequence $\{Y_k\}_{k=1}^\infty$ is a martingale with respect to $\{X_k\}_{k=1}^\infty$. Indeed, we have

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

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

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 $k \geq 1$,

$$\mathbb{E}[|D_k|] < \infty \quad \text{and} \quad \mathbb{E}[D_{k+1} \mid \mathcal{F}_k] = 0. \tag{2.26}$$

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 $k \geq 1$.

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

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 (2.25) 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

$$Y_n - Y_0 = \sum_{k=1}^{n} D_k, \tag{2.27}$$

where $\{D_k\}_{k=1}^{\infty}$ is a 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. 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 2.17, 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 martingale difference sequence, based on imposing a sub-exponential condition on the martingale differences.

Theorem 2.19 *Let $\{(D_k, \mathcal{F}_k)\}_{k=1}^{\infty}$ be a martingale difference sequence, and suppose that $\mathbb{E}[e^{\lambda D_k} \mid \mathcal{F}_{k-1}] \leq e^{\lambda^2 \nu_k^2/2}$ almost surely for any $|\lambda| < 1/\alpha_k$. Then the following hold:*

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

(b) *The sum satisfies the concentration inequality*

$$\mathbb{P}\left[\left|\sum_{k=1}^{n} D_k\right| \geq t\right] \leq \begin{cases} 2e^{-\frac{t^2}{2\sum_{k=1}^{n} \nu_k^2}} & \text{if } 0 \leq t \leq \frac{\sum_{k=1}^{n} \nu_k^2}{\alpha_*}, \\ 2e^{-\frac{t}{2\alpha_*}} & \text{if } t > \frac{\sum_{k=1}^{n} \nu_k^2}{\alpha_*}. \end{cases} \tag{2.28}$$

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. For any scalar λ such that $|\lambda| < \frac{1}{\alpha_*}$, condi-

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

tioning on $\mathcal{F}_{n-1}$ and applying iterated expectation yields

$$\begin{aligned}\mathbb{E}[e^{\lambda\left(\sum_{k=1}^{n} D_k\right)}] &= \mathbb{E}[e^{\lambda\left(\sum_{k=1}^{n-1} D_k\right)}\mathbb{E}[e^{\lambda D_n} \mid \mathcal{F}_{n-1}]] \\ &\leq \mathbb{E}[e^{\lambda \sum_{k=1}^{n-1} D_k}]e^{\lambda^2 \nu_n^2/2}, \end{aligned} \tag{2.29}$$

where the inequality follows from the stated assumption on D_n. Iterating this procedure yields the bound $\mathbb{E}[e^{\lambda \sum_{k=1}^{n} D_k}] \leq e^{\lambda^2 \sum_{k=1}^{n} \nu_k^2/2}$, valid for all $|\lambda| < \frac{1}{\alpha_*}$. By definition, we conclude that $\sum_{k=1}^{n} D_k$ is sub-exponential with parameters $\left(\sqrt{\sum_{k=1}^{n} \nu_k^2}, \alpha_*\right)$, as claimed. The tail bound (2.28) follows by applying Proposition 2.9. □

In order for Theorem 2.19 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 $\alpha = 0$). As discussed previously, bounded random variables are sub-Gaussian, which leads to the following corollary:

Corollary 2.20 (Azuma–Hoeffding) *Let $(\{(D_k, \mathcal{F}_k)\}_{k=1}^{\infty})$ be a martingale difference sequence for which there are constants $\{(a_k, b_k)\}_{k=1}^{n}$ such that $D_k \in [a_k, b_k]$ almost surely for all $k = 1, \ldots, n$. Then, for all $t \geq 0$,*

$$\mathbb{P}\left[\left|\sum_{k=1}^{n} D_k\right| \geq t\right] \leq 2e^{-\frac{2t^2}{\sum_{k=1}^{n}(b_k - a_k)^2}}. \tag{2.30}$$

Proof Recall the decomposition (2.29) in the proof of Theorem 2.19; from the structure of this argument, it suffices to show that $\mathbb{E}[e^{\lambda D_k} \mid \mathcal{F}_{k-1}] \leq e^{\lambda^2 (b_k - a_k)^2/8}$ almost surely for each $k = 1, 2, \ldots, n$. But since $D_k \in [a_k, b_k]$ almost surely, the conditioned variable $(D_k \mid \mathcal{F}_{k-1})$ also belongs to this interval almost surely, and hence from the result of Exercise 2.4, it is sub-Gaussian with parameter at most $\sigma = (b_k - a_k)/2$. □

An important application of Corollary 2.20 concerns functions that satisfy a bounded difference property. Let us first introduce some convenient notation. Given vectors $x, x' \in \mathbb{R}^n$ and an index $k \in \{1, 2, \ldots, n\}$, we define a new vector $x^{\backslash k} \in \mathbb{R}^n$ via

$$x_j^{\backslash k} := \begin{cases} x_j & \text{if } j \neq k, \\ x'_k & \text{if } j = k. \end{cases} \tag{2.31}$$

With this notation, we say that $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the bounded difference inequality with parameters $(L_1, \ldots, L_n)$ if, for each index $k = 1, 2, \ldots, n$,

$$|f(x) - f(x^{\backslash k})| \leq L_k \qquad \text{for all } x, x' \in \mathbb{R}^n. \tag{2.32}$$

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 coordinates.

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

Corollary 2.21 (Bounded differences inequality) *Suppose that f satisfies the bounded difference property* (2.32) *with parameters $(L_1, \ldots, L_n)$ and that the random vector $X = (X_1, X_2, \ldots, X_n)$ has independent components. Then*

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2e^{-\frac{2t^2}{\sum_{k=1}^n L_k^2}} \qquad \text{for all } t \geq 0. \tag{2.33}$$

Proof Recalling the Doob martingale introduced in Example 2.17, consider the associated martingale difference sequence

$$D_k = \mathbb{E}[f(X) \mid X_1, \ldots, X_k] - \mathbb{E}[f(X) \mid X_1, \ldots, X_{k-1}]. \tag{2.34}$$

We claim that D_k lies in an interval of length at most L_k almost surely. In order to prove this claim, define the random variables

$$\begin{aligned} A_k &:= \inf_x \mathbb{E}[f(X) \mid X_1, \ldots, X_{k-1}, x] - \mathbb{E}[f(X) \mid X_1, \ldots, X_{k-1}] \\ \text{and} \qquad B_k &:= \sup_x \mathbb{E}[f(X) \mid X_1, \ldots, X_{k-1}, x] - \mathbb{E}[f(X) \mid X_1, \ldots, X_{k-1}]. \end{aligned}$$

On one hand, we have

$$D_k - A_k = \mathbb{E}[f(X) \mid X_1, \ldots, X_k] - \inf_x \mathbb{E}[f(X) \mid X_1, \ldots, X_{k-1}, x],$$

so that $D_k \geq A_k$ almost surely. A similar argument shows that $D_k \leq B_k$ almost surely.

We now need to show that $B_k - A_k \leq L_k$ almost surely. Observe that by the independence of $\{X_k\}_{k=1}^n$, we have

$$\mathbb{E}[f(X) \mid x_1, \ldots, x_k] = \mathbb{E}_{k+1}[f(x_1, \ldots, x_k, X_{k+1}^n)] \qquad \text{for any vector } (x_1, \ldots, x_k),$$

where $\mathbb{E}_{k+1}$ denotes expectation over $X_{k+1}^n := (X_{k+1}, \ldots, X_n)$. Consequently, we have

$$\begin{aligned} B_k - A_k &= \sup_x \mathbb{E}_{k+1}[f(X_1, \ldots, X_{k-1}, x, X_{k+1}^n] - \inf_x \mathbb{E}_{k+1}[f(X_1, \ldots, X_{k-1}, x, X_{k+1}^n] \\ &\leq \sup_{x,y} \left| \mathbb{E}_{k+1}[f(X_1, \ldots, X_{k-1}, x, X_{k+1}^n) - f(X_1, \ldots, X_{k-1}, y, X_{k+1}^n)] \right| \\ &\leq L_k, \end{aligned}$$

using the bounded differences assumption. Thus, the variable D_k lies within an interval of length L_k at most surely, so that the claim follows as a corollary of the Azuma–Hoeffding inequality. □

Remark: In the special case when f is L-Lipschitz with respect to the Hamming norm, Corollary 2.21 implies that

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2e^{-\frac{2t^2}{nL^2}} \qquad \text{for all } t \geq 0. \tag{2.35}$$

Let us consider some examples to illustrate.

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

Example 2.22 (Classical Hoeffding from bounded differences) As a warm-up, let us show how the classical Hoeffding bound (2.11) for bounded variables—say $X_i \in [a, b]$ almost surely—follows as an immediate corollary of the bound (2.35). Consider the function $f(x_1, \ldots, x_n) = \sum_{i=1}^n (x_i - \mu_i)$, where $\mu_i = \mathbb{E}[X_i]$ is the mean of the ith random variable. For any index $k \in \{1, \ldots, n\}$, we have

$$\begin{aligned} |f(x) - f(x^{\backslash k})| &= |(x_k - \mu_k) - (x'_k - \mu_k)| \\ &= |x_k - x'_k| \leq b - a, \end{aligned}$$

showing that f satisfies the bounded difference inequality in each coordinate with parameter $L = b - a$. Consequently, it follows from the bounded difference inequality (2.35) that

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

which is the classical Hoeffding bound for independent random variables. ♣

The class of U-statistics frequently arise in statistical problems; let us now study their concentration properties.

Example 2.23 (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

$$U := \frac{1}{\binom{n}{2}} \sum_{j<k} g(X_j, X_k) \tag{2.36}$$

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 pairwise 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 fact can be revealed by a martingale analysis. If g is bounded (say $\|g\|_\infty \leq b$), then Corollary 2.21 can be used to establish the concentration of U around its mean. Viewing U as a function $f(x) = f(x_1, \ldots, x_n)$, for any given coordinate k, we have

$$\begin{aligned} |f(x) - f(x^{\backslash k})| &\leq \frac{1}{\binom{n}{2}} \sum_{j \neq k} |g(x_j, x_k) - g(x_j, x'_k)| \\ &\leq \frac{(n-1)(2b)}{\binom{n}{2}} = \frac{4b}{n}, \end{aligned}$$

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

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

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. ♣

Martingales and the bounded difference property also play an important role in analyzing the properties of random graphs, and other random combinatorial structures.

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

Example 2.24 (Clique number in random graphs) An undirected graph is a pair $G = (V, E)$, composed of a vertex set $V = \{1, \ldots, d\}$ and an edge set E, where each edge $e = (i, j)$ is an unordered pair of distinct vertices ($i \neq j$). A graph clique C is a subset of vertices such that $(i, j) \in E$ for all $i, j \in C$. The clique number $C(G)$ of the graph is the cardinality of the largest clique—note that $C(G) \in [1, d]$. When the edges E of the graph are drawn according to some random process, then the clique number $C(G)$ is a random variable, and we can study its concentration around its mean $\mathbb{E}[C(G)]$.

The *Erdös–Rényi* ensemble of random graphs is one of the most well-studied models: it is defined by a parameter $p \in (0, 1)$ that specifies the probability with which each edge (i, j) is included in the graph, independently across all $\binom{d}{2}$ edges. More formally, for each $i < j$, let us introduce a Bernoulli *edge-indicator variable* X_{ij} with parameter p, where $X_{ij} = 1$ means that edge (i, j) is included in the graph, and $X_{ij} = 0$ means that it is not included.

Note that the $\binom{d}{2}$-dimensional random vector $Z := \{X_{ij}\}_{i<j}$ specifies the edge set; thus, we may view the clique number $C(G)$ as a function $Z \mapsto f(Z)$. Let Z' denote a vector in which a single coordinate of Z has been changed, and let G' and G be the associated graphs. It is easy to see that $C(G')$ can differ from $C(G)$ by at most 1, so that $|f(Z') - f(Z)| \leq 1$. Thus, the function $C(G) = f(Z)$ satisfies the bounded difference property in each coordinate with parameter $L = 1$, so that

$$\mathbb{P}[\tfrac{1}{n}|C(G) - \mathbb{E}[C(G)]| \geq \delta] \leq 2e^{-2n\delta^2}.$$

Consequently, we see that the clique number of an Erdös–Rényi random graph is very sharply concentrated around its expectation. ♣

Finally, let us study concentration of the Rademacher complexity, a notion that plays a central role in our subsequent development in Chapters 4 and 5.

Example 2.25 (Rademacher complexity) Let $\{\varepsilon_k\}_{k=1}^n$ be an i.i.d. sequence of Rademacher variables (i.e., taking the values $\{-1, +1\}$ equiprobably, as in Example 2.3). Given a collection of vectors $\mathcal{A} \subset \mathbb{R}^n$, define the random variable[1]

$$Z := \sup_{a \in \mathcal{A}} \left[\sum_{k=1}^n a_k \varepsilon_k \right] = \sup_{a \in \mathcal{A}} [\langle a, \varepsilon \rangle]. \tag{2.37}$$

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 $\mathcal{A}$.

Let us now show how Corollary 2.21 can be used to establish that $Z(\mathcal{A})$ is sub-Gaussian. Viewing $Z(\mathcal{A})$ as a function $(\varepsilon_1, \ldots, \varepsilon_n) \mapsto f(\varepsilon_1, \ldots, \varepsilon_n)$, we need to bound the maximum change when coordinate k is changed. Given two Rademacher vectors $\varepsilon, \varepsilon' \in \{-1, +1\}^n$, recall our definition (2.31) of the modified vector $\varepsilon^{\backslash k}$. Since $f(\varepsilon^{\backslash k}) \geq \langle a, \varepsilon^{\backslash k} \rangle$ for any $a \in \mathcal{A}$, we have

$$\langle a, \varepsilon \rangle - f(\varepsilon^{\backslash k}) \leq \langle a, \varepsilon - \varepsilon^{\backslash k} \rangle = a_k(\varepsilon_k - \varepsilon'_k) \leq 2|a_k|.$$

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

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

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

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

Since the same argument applies with the roles of ε and $\varepsilon^{\backslash k}$ reversed, we conclude that f satisfies the bounded difference inequality in coordinate k with parameter $2\sup_{a\in\mathcal{A}}|a_k|$. Consequently, Corollary 2.21 implies that the random variable $Z(\mathcal{A})$ is sub-Gaussian with parameter at most $2\sqrt{\sum_{k=1}^n \sup_{a\in\mathcal{A}} a_k^2}$. This sub-Gaussian parameter can be reduced to the (potentially much) smaller quantity $\sqrt{\sup_{a\in\mathcal{A}}\sum_{k=1}^n a_k^2}$ using alternative techniques; in particular, see Example 3.5 in Chapter 3 for further details. ♣

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 a function $f\colon \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz with respect to the Euclidean norm $\|\cdot\|_2$ if

$$|f(x) - f(y)| \leq L\|x - y\|_2 \qquad \text{for all } x, y \in \mathbb{R}^n. \tag{2.38}$$

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

Theorem 2.26 *Let $(X_1, \ldots, X_n)$ be a vector of i.i.d. standard Gaussian variables, and let $f\colon \mathbb{R}^n \to \mathbb{R}$ 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*

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2e^{-\frac{t^2}{2L^2}} \qquad \text{for all } t \geq 0. \tag{2.39}$$

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 (2.39) 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. For a differentiable function, the Lipschitz property guarantees that $\|\nabla f(x)\|_2 \leq L$ for all $x \in \mathbb{R}^n$. 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.

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

Lemma 2.27 *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable. Then for any convex function $\phi : \mathbb{R} \to \mathbb{R}$, we have*

$$\mathbb{E}[\phi(f(X) - \mathbb{E}[f(X)])] \leq \mathbb{E}\Big[\phi\Big(\frac{\pi}{2}\langle \nabla f(X),\, Y\rangle\Big)\Big], \tag{2.40}$$

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

We now prove the theorem using this lemma. For any fixed $\lambda \in \mathbb{R}$, applying inequality (2.40) to the convex function $t \mapsto e^{\lambda t}$ yields

$$\begin{aligned}\mathbb{E}_X\Big[\exp\Big(\lambda\{f(X) - \mathbb{E}[f(X)]\}\Big)\Big] &\leq \mathbb{E}_{X,Y}\Big[\exp\Big(\frac{\lambda\pi}{2}\langle Y,\, \nabla f(X)\rangle\Big)\Big] \\ &= \mathbb{E}_X\Big[\exp\Big(\frac{\lambda^2\pi^2}{8}\|\nabla f(X)\|_2^2\Big)\Big],\end{aligned}$$

where we have used the independence of X and Y to first take the expectation over Y marginally, and the fact that $\langle Y, \nabla f(x)\rangle$ is a zero-mean Gaussian variable with variance $\|\nabla f(x)\|_2^2$. Due to the Lipschitz condition on f, we have $\|\nabla f(x)\|_2 \leq L$ for all $x \in \mathbb{R}^n$, whence

$$\mathbb{E}\Big[\exp\Big(\lambda\{f(X) - \mathbb{E}[f(X)]\}\Big)\Big] \leq e^{\frac{1}{8}\lambda^2\pi^2L^2},$$

which shows that $f(X) - \mathbb{E}[f(X)]$ is sub-Gaussian with parameter at most $\frac{\pi L}{2}$. The tail bound

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

follows from Proposition 2.5.

It remains to prove Lemma 2.27, and we do so via a classical 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

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

By the convexity of ϕ, we have

$$\mathbb{E}_X[\phi(f(X) - \mathbb{E}_Y[f(Y)])] \leq \mathbb{E}_{X,Y}[\phi(f(X) - f(Y))]. \tag{2.41}$$

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

$$f(X) - f(Y) = \int_0^{\pi/2} \frac{d}{d\theta} f(Z(\theta))\, d\theta = \int_0^{\pi/2} \langle \nabla f(Z(\theta)),\, Z'(\theta)\rangle\, d\theta, \tag{2.42}$$

where $Z'(\theta) \in \mathbb{R}^n$ denotes the elementwise derivative, a vector with the components $Z_k'(\theta) = X_k\cos\theta - Y_k\sin\theta$. Substituting the integral representation (2.42) into our earlier bound (2.41)

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

yields

$$\begin{aligned}\mathbb{E}_X[\phi(f(X) - \mathbb{E}_Y[f(Y)])] &\leq \mathbb{E}_{X,Y}\left[\phi\left(\int_0^{\pi/2} \langle \nabla f(Z(\theta)),\, Z'(\theta)\rangle \, d\theta\right)\right] \\ &= \mathbb{E}_{X,Y}\left[\phi\left(\frac{1}{\pi/2}\int_0^{\pi/2} \frac{\pi}{2}\langle \nabla f(Z(\theta)),\, Z'(\theta)\rangle \, d\theta\right)\right] \\ &\leq \frac{1}{\pi/2}\int_0^{\pi/2} \mathbb{E}_{X,Y}\left[\phi\left(\frac{\pi}{2}\langle \nabla f(Z(\theta)),\, Z'(\theta)\rangle\right)\right] d\theta, \qquad (2.43)\end{aligned}$$

where the final step again uses convexity of ϕ. By the rotation invariance of the Gaussian distribution, for each $\theta \in [0, \pi/2]$, the pair $(Z_k(\theta), Z_k'(\theta))$ is a jointly Gaussian vector, with zero mean and identity covariance $\mathbf{I}_2$. Therefore, the expectation inside the integral in equation (2.43) does not depend on θ, and hence

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

where $(\widetilde{X}, \widetilde{Y})$ are independent standard Gaussian n-vectors. This completes the proof of the bound (2.40). □

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). However, 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 2.26 is useful for a broad range of problems; let us consider some examples to illustrate.

Example 2.28 (χ^2 concentration) For a given sequence $\{Z_k\}_{k=1}^n$ of i.i.d. standard normal variates, the random variable $Y := \sum_{k=1}^n Z_k^2$ follows a χ^2-distribution with n degrees of freedom. The most direct way to obtain tail bounds on Y is by noting that Z_k^2 is sub-exponential, and exploiting independence (see Example 2.11). 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 2.26 implies that

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

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

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

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

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

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

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

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

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 2.11. ♣

Example 2.29 (Order statistics) Given a random vector $(X_1, X_2, \ldots, X_n)$, its order statistics are obtained by reordering its entries in a non-decreasing manner—namely as

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n-1)} \leq X_{(n)}. \tag{2.45}$$

As particular cases, we have $X_{(n)} = \max_{k=1,\ldots,n} X_k$ and $X_{(1)} = \min_{k=1,\ldots,n} X_k$. 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 2.26 implies that

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

Example 2.30 (Gaussian complexity) This example is closely related to our earlier discussion of Rademacher complexity in Example 2.25. 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]

$$Z := \sup_{a \in \mathcal{A}} \left[\sum_{k=1}^n a_k W_k \right] = \sup_{a \in \mathcal{A}} \langle a, W \rangle. \tag{2.46}$$

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 $(w_1, \ldots, w_n) \mapsto f(w_1, \ldots, w_n)$, 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 2.25, we have the upper bound

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

where $D(\mathcal{A}) = \sup_{a \in \mathcal{A}} \|a\|_2$ is the Euclidean width of the set. The same argument holds with the roles of w and w' reversed, and hence

$$|f(w) - f(w')| \leq D(\mathcal{A}) \, \|w - w'\|_2.$$

Consequently, Theorem 2.26 implies that

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

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

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

Example 2.31 (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

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

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

Conditioned on $\widetilde{w}$, the variable Z is a zero-mean Gaussian variable with variance $\|\mathbf{Q}\widetilde{w}\|_2^2 = \widetilde{w}^{\mathrm{T}} \mathbf{Q}^2 \widetilde{w}$, whence

$$\mathbb{P}[|Z| \geq \delta \mid \widetilde{w}] \leq 2e^{-\frac{\delta^2}{2\|\mathbf{Q}\widetilde{w}\|_2^2}}. \tag{2.48}$$

Let us now control the random variable $Y := \|\mathbf{Q}\widetilde{w}\|_2$. Viewed as a function of the Gaussian vector $\widetilde{w}$, it is Lipschitz with constant

$$|||\mathbf{Q}|||_2 := \sup_{\|u\|_2 = 1} \|\mathbf{Q}u\|_2, \tag{2.49}$$

corresponding to the *ℓ_2-operator norm* of the matrix $\mathbf{Q}$. Moreover, by Jensen's inequality, we have $\mathbb{E}[Y] \leq \sqrt{\mathbb{E}[\widetilde{w}^{\mathrm{T}} \mathbf{Q}^2 \widetilde{w}]} = |||\mathbf{Q}|||_{\mathrm{F}}$, where

$$|||\mathbf{Q}|||_{\mathrm{F}} := \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} Q_{ij}^2} \tag{2.50}$$

is the *Frobenius norm* of the matrix $\mathbf{Q}$. Putting together the pieces yields the tail bound

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

Note that $(|||\mathbf{Q}|||_{\mathrm{F}} + t)^2 \leq 2|||\mathbf{Q}|||_{\mathrm{F}}^2 + 2t^2$. Consequently, setting $t^2 = \delta |||\mathbf{Q}|||_2$ and simplifying yields

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

Putting together the pieces, we find that

$$\begin{aligned}\mathbb{P}[|Z| \geq \delta] &\leq 2 \exp\Big(-\frac{\delta^2}{4|||\mathbf{Q}|||_{\mathrm{F}}^2 + 4\delta |||\mathbf{Q}|||_2}\Big) + 2 \exp\Big(-\frac{\delta}{2|||\mathbf{Q}|||_2}\Big) \\ &\leq 4 \exp\Big(-\frac{\delta^2}{4|||\mathbf{Q}|||_{\mathrm{F}}^2 + 4\delta |||\mathbf{Q}|||_2}\Big).\end{aligned}$$

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

Example 2.32 (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

$$\sigma_1(\mathbf{X}) \geq \sigma_2(\mathbf{X}) \geq \cdots \geq \sigma_d(\mathbf{X}) \geq 0$$

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

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

$$\max_{k=1,\ldots,d} |\sigma_k(\mathbf{X}) - \sigma_k(\mathbf{Y})| \leq |||\mathbf{X} - \mathbf{Y}|||_2 \leq |||\mathbf{X} - \mathbf{Y}|||_{\mathrm{F}}, \tag{2.51}$$

where $|||\cdot|||_{\mathrm{F}}$ denotes the Frobenius norm. The inequality (2.51) shows that each singular value $\sigma_k(\mathbf{X})$ is a 1-Lipschitz function of the random matrix, so that Theorem 2.26 implies that, for each $k = 1, \ldots, d$, we have

$$\mathbb{P}[|\sigma_k(\mathbf{X}) - \mathbb{E}[\sigma_k(\mathbf{X})]| \geq \delta] \leq 2e^{-\frac{\delta^2}{2}} \qquad \text{for all } \delta \geq 0. \tag{2.52}$$

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. ♣

2.4 Appendix A: Equivalent versions of sub-Gaussian variables

In this appendix, we prove Theorem 2.6. We establish the equivalence by proving the circle of implications (I) $\Rightarrow$ (II) $\Rightarrow$ (III) $\Rightarrow$ (I), followed by the equivalence (I) $\Leftrightarrow$ (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 Mills ratio for Gaussian tails, if $Z \sim \mathcal{N}(0, 2\sigma^2)$, then we have

$$\mathbb{P}[Z \geq t] \geq \left(\frac{\sqrt{2}\sigma}{t} - \frac{(\sqrt{2}\sigma)^3}{t^3}\right) e^{-\frac{t^2}{4\sigma^2}} \qquad \text{for all } t > 0. \tag{2.53}$$

(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,

$$\mathbb{P}[Z \geq t] \geq \mathbb{P}[Z \geq 2\sigma] \geq \left(\frac{1}{\sqrt{2}} - \frac{1}{2\sqrt{2}}\right) e^{-1} = \frac{1}{\sqrt{8}e}.$$

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 Mills ratio (2.53) and the sub-Gaussian tail bound and making the substitution $s = t/\sigma$, we find

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

that

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

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 Z are given by

$$\mathbb{E}[Z^{2k}] = \frac{(2k)!}{2^k k!}\tau^{2k}, \qquad \text{for } k = 1, 2, \ldots, \tag{2.54}$$

whence

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

Consequently, the moment bound (2.12c) holds with $\theta = c\tau$.

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

$$\mathbb{E}[e^{\lambda X}] \leq 1 + \sum_{k=2}^{\infty} \frac{|\lambda|^k \mathbb{E}[|X|^k]}{k!}, \tag{2.55}$$

where we have used the fact $\mathbb{E}[X] = 0$ to eliminate 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

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

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

$$\mathbb{E}[|\lambda X|^{2k+1}] \overset{\text{(i)}}{\leq} (\mathbb{E}[|\lambda X|^{2k}]\mathbb{E}[|\lambda X|^{2k+2}])^{1/2} \overset{\text{(ii)}}{\leq} \tfrac{1}{2}(\lambda^{2k}\mathbb{E}[X^{2k}] + \lambda^{2k+2}\mathbb{E}[X^{2k+2}]), \tag{2.56}$$

where step (i) follows from the Cauchy–Schwarz inequality; and step (ii) follows from

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

the arithmetic–geometric mean inequality. Applying this bound to the power-series expansion (2.55), we obtain

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

which establishes the claim.

Implication (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

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

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

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

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

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

Taking expectations with respect to X, we conclude that

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

which establishes the claim.

Implication (IV) $\Rightarrow$ (I): Applying the bound $e^u \leq u + e^{9u^2/16}$ with $u = \lambda X$ and then taking expectations, we find that

$$\mathbb{E}[e^{\lambda X}] \leq \mathbb{E}[\lambda X] + \mathbb{E}[e^{\frac{9\lambda^2X^2}{16}}] = \mathbb{E}[e^{\frac{sX^2}{2\sigma^2}}] \leq \frac{1}{\sqrt{1-s}},$$

valid whenever $s = \frac{9}{8}\lambda^2\sigma^2$ is strictly less than 1. Noting that $\frac{1}{\sqrt{1-s}} \leq e^s$ for all $s \in [0, \frac{1}{2}]$ and that $s < \frac{1}{2}$ whenever $|\lambda| < \frac{2}{3\sigma}$, we conclude that

$$\mathbb{E}[e^{\lambda X}] \leq e^{\frac{9}{8}\lambda^2\sigma^2} \qquad \text{for all } |\lambda| < \tfrac{2}{3\sigma}. \tag{2.57a}$$

It remains to establish a similar upper bound for $|\lambda| \geq \frac{2}{3\sigma}$. Note that, for any $\alpha > 0$, the functions $f(u) = \frac{u^2}{2\alpha}$ and $f^*(v) = \frac{\alpha v^2}{2}$ are conjugate duals. Thus, the Fenchel–Young

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

inequality implies that $uv \leq \frac{u^2}{2\alpha} + \frac{\alpha v^2}{2}$, valid for all $u, v \in \mathbb{R}$ and $\alpha > 0$. We apply this inequality with $u = \lambda$, $v = X$ and $\alpha = c/\sigma^2$ for a constant $c > 0$ to be chosen; doing so yields

$$\mathbb{E}[e^{\lambda X}] \leq \mathbb{E}[e^{\frac{\lambda^2\sigma^2}{2c} + \frac{cX^2}{2\sigma^2}}] = e^{\frac{\lambda^2\sigma^2}{2c}} \mathbb{E}[e^{\frac{cX^2}{2\sigma^2}}] \overset{\text{(ii)}}{\leq} e^{\frac{\lambda^2\sigma^2}{2c}} e^c,$$

where step (ii) is valid for any $c \in (0, 1/2)$, using the same argument that led to the bound (2.57a). In particular, setting $c = 1/4$ yields $\mathbb{E}[e^{\lambda X}] \leq e^{2\lambda^2\sigma^2} e^{1/4}$.

Finally, when $|\lambda| \geq \frac{2}{3\sigma}$, then we have $\frac{1}{4} \leq \frac{9}{16}\lambda^2\sigma^2$, and hence

$$\mathbb{E}[e^{\lambda X}] \leq e^{2\lambda^2\sigma^2 + \frac{9}{16}\lambda^2\sigma^2} \leq e^{3\lambda^2\sigma^2}. \tag{2.57b}$$

This inequality, combined with the bound (2.57a), completes the proof.

2.5 Appendix B: Equivalent versions of sub-exponential variables

This appendix is devoted to the proof of Theorem 2.13. In particular, we prove the chain of equivalences I $\Leftrightarrow$ II $\Leftrightarrow$ III, followed by the equivalence II $\Leftrightarrow$ IV.

(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

$$\mathbb{E}[e^{a|X|} \mathbb{I}[e^{a|X|} \leq e^{aT}]] \leq \int_0^{e^{aT}} \mathbb{P}[e^{a|X|} \geq t]\, dt \leq 1 + \int_1^{e^{aT}} \mathbb{P}\Big[|X| \geq \frac{\log t}{a}\Big] dt.$$

Applying the assumed tail bound, we obtain

$$\mathbb{E}[e^{a|X|} \mathbb{I}[e^{a|X|} \leq e^{aT}]] \leq 1 + c_1 \int_1^{e^{aT}} e^{-\frac{c_2 \log t}{a}}\, dt = 1 + c_1 \int_1^{e^{aT}} t^{-c_2/a}\, dt.$$

Thus, for any $a \in [0, \frac{c_2}{2}]$, we have

$$\mathbb{E}[e^{a|X|} \mathbb{I}[e^{a|X|} \leq e^{aT}]] \leq 1 + \frac{c_1}{2}(1 - e^{-aT}) \leq 1 + \frac{c_1}{2}.$$

By taking the limit as $T \to \infty$, we conclude 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 $e^{|a||X|}$, it follows that $\mathbb{E}[e^{aX}]$ is finite for all $|a| \leq \frac{c_2}{2}$.

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

$$\mathbb{P}[X \geq t] \leq \mathbb{E}[e^{\frac{c_0 X}{2}}]\, e^{-\frac{c_0 t}{2}}.$$

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

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

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

$$\mathbb{E}[e^{\lambda X}] = 1 + \sum_{k=2}^{\infty} \frac{\lambda^k \mathbb{E}[X^k]}{k!} \qquad \text{for all } |\lambda| < a. \tag{2.58}$$

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

2.6 Bibliographic details and background

Further background and details on tail bounds can be found in various books (e.g., Saulis and Statulevicius, 1991; Petrov, 1995; Buldygin and Kozachenko, 2000; Boucheron et al., 2013). Classic papers on tail bounds include those of Bernstein (1937), Chernoff (1952), Bahadur and Ranga Rao (1960), Bennett (1962), Hoeffding (1963) and Azuma (1967). The idea of using the cumulant function to bound the tails of a random variable was first introduced by Bernstein (1937), and further developed by Chernoff (1952), whose name is now frequently associated with the method. The book by Saulis and Statulevicius (1991) provides a number of more refined results that can be established using cumulant-based techniques. The original work of Hoeffding (1963) 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 (1967) applies to general martingales that are sub-Gaussian in a conditional sense, as in Theorem 2.19.

The book by Buldygin and Kozachenko (2000) provides a range of results on sub-Gaussian and sub-exponential variates. In particular, Theorems 2.6 and 2.13 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 5.6 and the books (Ledoux and Talagrand, 1991; Buldygin and Kozachenko, 2000) for further background on Orlicz norms.

The Johnson–Lindenstrauss lemma, discussed in Example 2.12, was originally proved by Johnson and Lindenstrauss (1984) as an intermediate step in a more general result about Lipschitz embeddings. The original proof of the 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 sources (Vempala, 2004; Mahoney, 2011; Cormode, 2012; Kane and Nelson, 2014; Woodruff, 2014; Bourgain et al., 2015; Pilanci and Wainwright, 2015) for further details.

Tail bounds for U-statistics, as sketched out in Example 2.23, were derived by Hoeffding (1963). The book by de la Peña and Giné (1999) provides more advanced results, including extensions to uniform laws for U-processes and decoupling results. The bounded differences inequality (Corollary 2.21) 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 (1989), and the book by Boucheron et al. (2013).

Milman and Schechtman (1986) provide the short proof of Gaussian concentration for

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

Lipschitz functions, on which Theorem 2.26 is based. Ledoux (2001) provides an example of a Lipschitz function of an i.i.d. sequence of Rademacher variables (i.e., taking values $\{-1, +1\}$ equiprobably) for which sub-Gaussian concentration fails to hold (cf. p. 128 in his book). 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 nonparametric 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 (Ledoux and Talagrand, 1991; Yurinsky, 1995) 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 (Hanson and Wright, 1971; Wright, 1973); see the papers (Hsu et al., 2012b; Rudelson and Vershynin, 2013) for more modern treatments. The moment-based tail bound from Exercise 2.20 relies on a classical inequality due to Rosenthal (1970). 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 (1991) for further reading. The Ising model (2.74) discussed in Exercise 2.22 has a lengthy history dating back to Ising (1925). The book by Talagrand (2003) contains a wealth of information on spin glass models and their mathematical properties.

2.7 Exercises

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

(a) Provide a non-negative random variable X 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 (Mills ratio) 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

$$\phi(z)\left(\frac{1}{z} - \frac{1}{z^3}\right) \leq \mathbb{P}[Z \geq z] \leq \phi(z)\left(\frac{1}{z} - \frac{1}{z^3} + \frac{3}{z^5}\right) \qquad \text{for all } z > 0. \tag{2.59}$$

Exercise 2.3 (Polynomial Markov versus Chernoff) 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

$$\inf_{k=0,1,2,\ldots} \frac{\mathbb{E}[|X|^k]}{\delta^k} \leq \inf_{\lambda > 0} \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda \delta}}. \tag{2.60}$$

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

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

Exercise 2.4 (Sharp sub-Gaussian parameter for bounded random variable) Consider a random variable X with mean $\mu = \mathbb{E}[X]$, and such that, for some scalars $b > a$, $X \in [a, b]$ 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

$$\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\lambda^2\sigma^2}{2} + \lambda\mu} \qquad \text{for all } \lambda \in \mathbb{R}. \tag{2.61}$$

(a) Show that $\mathbb{E}[X] = \mu$.
(b) Show that $\text{var}(X) \leq \sigma^2$.
(c) Suppose that the smallest possible σ satisfying the inequality (2.61) is chosen. Is it then true that $\text{var}(X) = \sigma^2$? 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

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

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) This exercise is devoted to a proof of a strengthening of Bernstein's inequality, known as Bennett's inequality.

(a) Consider a zero-mean random variable such that $|X_i| \leq b$ for some $b > 0$. Prove that

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

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

$$\mathbb{P}\left[\sum_{i=1}^n X_i \geq n\delta\right] \leq \exp\left\{ -\frac{n\sigma^2}{b^2} h\left(\frac{b\delta}{\sigma^2}\right)\right\}, \tag{2.62}$$

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.

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

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

$$\mathbb{P}[Z \geq t] \leq C e^{-\frac{t^2}{2(\nu^2 + bt)}} \tag{2.63}$$

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 (2.15). Use part (a) to show that

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

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

$$D(\delta \| \alpha) := \delta \log \frac{\delta}{\alpha} + (1 - \delta) \log \frac{(1 - \delta)}{(1 - \alpha)} \tag{2.64}$$

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

(b) Show that

$$\frac{1}{n} \log \binom{n}{m} \geq \phi(\widetilde{\delta}) - \frac{\log(n + 1)}{n}, \tag{2.65a}$$

where $\phi(\widetilde{\delta}) = -\widetilde{\delta} \log(\widetilde{\delta}) - (1 - \widetilde{\delta}) \log(1 - \widetilde{\delta})$ is the binary entropy. (*Hint:* Let Y be a 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

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

where the Kullback–Leibler divergence $D(\delta \| \alpha)$ was previously defined (2.64).

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

Exercise 2.11 (Upper and lower bounds for 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

$$\mathbb{E}[Z_n] \leq \sqrt{2\sigma^2 \log n} + \frac{4\sigma}{\sqrt{2 \log n}} \qquad \text{for all } n \geq 2.$$

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

(b) Prove that

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

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

Exercise 2.12 (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.)

(a) Prove that

$$\mathbb{E}\left[\max_{i=1,\ldots,n} X_i\right] \leq \sqrt{2\sigma^2 \log n} \qquad \text{for all } n \geq 1. \tag{2.66}$$

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

(b) Prove that the random variable $Z = \max_{i=1,\ldots,n} |X_i|$ satisfies

$$\mathbb{E}[Z] \leq \sqrt{2\sigma^2 \log(2n)} \leq 2\sqrt{\sigma^2 \log n}, \tag{2.67}$$

valid for all $n \geq 2$.

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 the random variable $X_1 + X_2$ 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 $\sqrt{2}\sqrt{\sigma_1^2 + \sigma_2^2}$.

(c) In the same setting as part (b), show that $X_1 + X_2$ is sub-Gaussian with parameter at most $\sigma_1 + \sigma_2$.

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

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

$$\mathbb{P}[|X - \mathbb{E}[X]| \geq t] \leq c_1 e^{-c_2 t^2} \qquad \text{for all } t \geq 0. \tag{2.68}$$

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

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

(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 (2.68) holds, then for any median m_X, we have

$$\mathbb{P}[|X - m_X| \geq t] \leq c_3 e^{-c_4 t^2} \qquad \text{for all } t \geq 0, \tag{2.69}$$

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

(d) Conversely, show that whenever the median concentration bound (2.69) holds, then mean concentration (2.68) holds with $c_1 = 2c_3$ and $c_2 = \frac{c_4}{4}$.

Exercise 2.15 (Concentration and kernel density estimation) Let $\{X_i\}_{i=1}^n$ be an i.i.d. sequence of random variables drawn from a density f on the real line. A standard estimate of f is the *kernel density estimate*

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

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

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

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 $S_n = \left\| \sum_{i=1}^n X_i \right\|_{\mathbb{H}}$.

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

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

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

(*Note:* See Chapter 12 for basic background on Hilbert spaces.)

Exercise 2.17 (Hanson–Wright inequality) Given random variables $\{X_i\}_{i=1}^n$ and a positive semidefinite matrix $\mathbf{Q} \in \mathcal{S}_+^{n \times n}$, consider the random quadratic form

$$Z = \sum_{i=1}^n \sum_{j=1}^n \mathbf{Q}_{ij} X_i X_j. \tag{2.70}$$

The *Hanson–Wright inequality* guarantees that whenever the random variables $\{X_i\}_{i=1}^n$ are i.i.d. with mean zero, unit variance, and σ-sub-Gaussian, then there are universal constants (c_1, c_2) such that

$$\mathbb{P}[Z \geq \operatorname{trace}(\mathbf{Q}) + \sigma t] \leq 2 \exp\left\{ - \min\left(\frac{c_1 t}{|||\mathbf{Q}|||_2}, \frac{c_2 t^2}{|||\mathbf{Q}|||_{\mathrm{F}}^2} \right) \right\}, \tag{2.71}$$

where $|||\mathbf{Q}|||_2$ and $|||\mathbf{Q}|||_{\mathrm{F}}$ denote the operator and Frobenius norms, respectively. Prove this

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

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

$$\|X\|_\psi := \inf\{t > 0 \mid \mathbb{E}[\psi(t^{-1}|X|)] \leq 1\}, \tag{2.72}$$

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 $\|X\|_q = (\mathbb{E}[|X|^q])^{1/q}$. 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

$$\mathbb{P}[|X| > t] \leq c_1 \exp(-c_2 t^q) \qquad \text{for all } t > 0. \tag{2.73}$$

(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 (2.73). 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

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

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 $\|X_i\|_{2m} := (\mathbb{E}[X_i^{2m}])^{\frac{1}{2m}} \leq C_m$. Show that

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

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

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

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 collection of binary vectors, say $\{z^1, \ldots, z^N\}$, such that the *rescaled Hamming distortion*

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

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

is as small as possible. Each binary vector z^j is known as a codeword, and the full collection is called a codebook. 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 $\{z^1, \ldots, z^N\}$ with $N \leq 2^{nR}$ codewords, the probability of the event $\{d(X) \leq \delta\}$ goes to zero as n goes to infinity. (*Hint:* Let V^j be a {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 $\mathbb{P}[V \geq 1] \to +\infty$ as $n \to +\infty$. (*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

$$\mathbb{P}_\theta(x_1, \ldots, x_d) = \exp\left\{ \frac{1}{\sqrt{d}} \sum_{i \neq j} \theta_{jk} x_j x_k - F_d(\theta) \right\}, \tag{2.74}$$

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

$$F_d(\theta) := \log \left(\sum_{x \in \{-1,+1\}^d} \exp\left\{ \frac{1}{\sqrt{d}} \sum_{j \neq k} \theta_{jk} x_j x_k \right\} \right) \tag{2.75}$$

serves to normalize the distribution. The probability distribution (2.74) 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 (2.74) 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

$$\mathbb{P}\left[\frac{F_d(\theta)}{d} \geq \log 2 + \frac{\beta^2}{4} + t \right] \leq 2e^{-\beta d t^2/2} \qquad \text{for all } t > 0. \tag{2.76}$$

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press

Remark: Interestingly, it is known that, for any $\beta \in [0, 1)$, this upper tail bound captures the asymptotic behavior of $F_d(\theta)/d$ accurately, in that $\frac{F_d(\theta)}{d} \xrightarrow{a.s.} \log 2 + \beta^2/4$ as $d \to \infty$. By 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.

https://doi.org/10.1017/9781108627771.002 Published online by Cambridge University Press