3

Concentration of measure

Building upon the foundation of Chapter 2, this chapter is devoted to an exploration of more advanced material on the concentration of measure. In particular, our goal is to provide an overview of the different types of methods available to derive tail bounds and concentration inequalities. We begin in Section 3.1 with a discussion of the entropy method for concentration, and illustrate its use in deriving tail bounds for Lipschitz functions of independent random variables. In Section 3.2, we turn to some geometric aspects of concentration inequalities, a viewpoint that is historically among the oldest. Section 3.3 is devoted to the use of transportation cost inequalities for deriving concentration inequalities, a method that is in some sense dual to the entropy method, and well suited to certain types of dependent random variables. We conclude in Section 3.4 by deriving some tail bounds for empirical processes, including versions of the functional Hoeffding and Bernstein inequalities. These inequalities play an especially important role in our later treatment of nonparametric problems.

3.1 Concentration by entropic techniques

We begin our exploration with the entropy method and related techniques for deriving concentration inequalities.

3.1.1 Entropy and its properties

Given a convex function $\phi\colon \mathbb{R} \to \mathbb{R}$, it can be used to define a functional on the space of probability distributions via

$$\mathbb{H}_\phi(X) := \mathbb{E}[\phi(X)] - \phi(\mathbb{E}[X]),$$

where $X \sim \mathbb{P}$. This quantity, which is well defined for any random variable such that both X and $\phi(X)$ have finite expectations, is known as the *ϕ-entropy*[1] of the random variable X. By Jensen's inequality and the convexity of ϕ, the ϕ-entropy is always non-negative. As the name suggests, it serves as a measure of variability. For instance, in the most extreme case, we have $\mathbb{H}_\phi(X) = 0$ for any random variable such that X is equal to its expectation $\mathbb{P}$-almost-everywhere.

[1] The notation $\mathbb{H}_\phi(X)$ has the potential to mislead, since it suggests that the entropy is a function of X, and hence a random variable. To be clear, the entropy $\mathbb{H}_\phi$ is a functional that acts on the probability measure $\mathbb{P}$, as opposed to the random variable X.

https://doi.org/10.1017/9781108627771.003 Published online by Cambridge University Press

There are various types of entropies, depending on the choice of the underlying convex function ϕ. Some of these entropies are already familiar to us. For example, the convex function $\phi(u) = u^2$ yields

$$\mathbb{H}_\phi(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \text{var}(X),$$

corresponding to the usual variance of the random variable X. Another interesting choice is the convex function $\phi(u) = -\log u$ defined on the positive real line. When applied to the positive random variable $Z := e^{\lambda X}$, this choice of ϕ yields

$$\mathbb{H}_\phi(e^{\lambda X}) = -\lambda\mathbb{E}[X] + \log \mathbb{E}[e^{\lambda X}] = \log \mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}],$$

a type of entropy corresponding to the centered cumulant generating function. In Chapter 2, we have seen how both the variance and the cumulant generating function are useful objects for obtaining concentration inequalities—in particular, in the form of Chebyshev's inequality and the Chernoff bound, respectively.

Throughout the remainder of this chapter, we focus on a slightly different choice of entropy functional, namely the convex function $\phi\colon [0, \infty) \to \mathbb{R}$ defined as

$$\phi(u) := u \log u \quad \text{for } u > 0, \qquad \text{and} \quad \phi(0) := 0. \tag{3.1}$$

For any non-negative random variable $Z \geq 0$, it defines the ϕ-entropy given by

$$\mathbb{H}(Z) = \mathbb{E}[Z \log Z] - \mathbb{E}[Z] \log \mathbb{E}[Z], \tag{3.2}$$

assuming that all relevant expectations exist. In the remainder of this chapter, we omit the subscript ϕ, since the choice (3.1) is to be implicitly understood.

The reader familiar with information theory may observe that the entropy (3.2) is closely related to the Shannon entropy, as well as the Kullback–Leibler divergence; see Exercise 3.1 for an exploration of this connection. As will be clarified in the sequel, the most attractive property of the ϕ-entropy (3.2) is its so-called tensorization when applied to functions of independent random variables.

For the random variable $Z := e^{\lambda X}$, the entropy has an explicit expression as a function of the moment generating function $\varphi_x(\lambda) = \mathbb{E}[e^{\lambda X}]$ and its first derivative. In particular, a short calculation yields

$$\mathbb{H}(e^{\lambda X}) = \lambda\varphi'_x(\lambda) - \varphi_x(\lambda) \log \varphi_x(\lambda). \tag{3.3}$$

Consequently, if we know the moment generating function of X, then it is straightforward to compute the entropy $\mathbb{H}(e^{\lambda X})$. Let us consider a simple example to illustrate:

Example 3.1 (Entropy of a Gaussian random variable) For the scalar Gaussian variable $X \sim \mathcal{N}(0, \sigma^2)$, we have $\varphi_x(\lambda) = e^{\lambda^2\sigma^2/2}$. By taking derivatives, we find that $\varphi'_x(\lambda) = \lambda\sigma^2\varphi_x(\lambda)$, and hence

$$\mathbb{H}(e^{\lambda X}) = \lambda^2\sigma^2\varphi_x(\lambda) - \tfrac{1}{2}\lambda^2\sigma^2\, \varphi_x(\lambda) = \tfrac{1}{2}\lambda^2\sigma^2\, \varphi_x(\lambda). \tag{3.4}$$

♣

Given that the moment generating function can be used to obtain concentration inequalities via the Chernoff method, this connection suggests that there should also be a connection between the entropy (3.3) and tail bounds. It is the goal of the following sections to make

https://doi.org/10.1017/9781108627771.003 Published online by Cambridge University Press

this connection precise for various classes of random variables. We then show how the entropy based on $\phi(u) = u \log u$ has a certain tensorization property that makes it particularly well suited to dealing with general Lipschitz functions of collections of random variables.

3.1.2 Herbst argument and its extensions

Intuitively, the entropy is a measure of the fluctuations in a random variable, so that control on the entropy should translate into bounds on its tails. The Herbst argument makes this intuition precise for a certain class of random variables. In particular, suppose that there is a constant $\sigma > 0$ such that the entropy of $e^{\lambda X}$ satisfies an upper bound of the form

$$\mathbb{H}(e^{\lambda X}) \leq \tfrac{1}{2}\sigma^2 \lambda^2 \, \varphi_{\mathrm{x}}(\lambda). \tag{3.5}$$

Note that by our earlier calculation in Example 3.1, any Gaussian variable $X \sim \mathcal{N}(0, \sigma^2)$ satisfies this condition *with equality* for all $\lambda \in \mathbb{R}$. Moreover, as shown in Exercise 3.7, any bounded random variable satisfies an inequality of the form (3.5).

Of interest here is the other implication: What does the entropy bound (3.5) imply about the tail behavior of the random variable? The classical Herbst argument answers this question, in particular showing that any such variable must have sub-Gaussian tail behavior.

Proposition 3.2 (Herbst argument) *Suppose that the entropy* $\mathbb{H}(e^{\lambda X})$ *satisfies inequality* (3.5) *for all* $\lambda \in I$, *where* I *can be either of the intervals* $[0, \infty)$ *or* $\mathbb{R}$. *Then* X *satisfies the bound*

$$\log \mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}] \leq \tfrac{1}{2}\lambda^2\sigma^2 \qquad \textit{for all } \lambda \in I. \tag{3.6}$$

Remarks: When $I = \mathbb{R}$, then the inequality (3.6) is equivalent to asserting that the centered variable $X - \mathbb{E}[X]$ is sub-Gaussian with parameter σ. Via an application of the usual Chernoff argument, the bound (3.6) with $I = [0, \infty)$ implies the one-sided tail bound

$$\mathbb{P}[X \geq \mathbb{E}[X] + t] \leq e^{-\frac{t^2}{2\sigma^2}}, \tag{3.7}$$

and with $I = \mathbb{R}$, it implies the two-sided bound $\mathbb{P}[|X - \mathbb{E}[X]| \geq t] \leq 2e^{-\frac{t^2}{2\sigma^2}}$. Of course, these are the familiar tail bounds for sub-Gaussian variables discussed previously in Chapter 2.

Proof Recall the representation (3.3) of entropy in terms of the moment generating function. Combined with the assumed upper bound (3.5), we conclude that the moment generating function $\varphi \equiv \varphi_{\mathrm{x}}$ satisfies the differential inequality

$$\lambda\varphi'(\lambda) - \varphi(\lambda)\log\varphi(\lambda) \leq \tfrac{1}{2}\sigma^2\lambda^2\,\varphi(\lambda), \qquad \text{valid for all } \lambda \geq 0. \tag{3.8}$$

Define the function $G(\lambda) = \frac{1}{\lambda}\log\varphi(\lambda)$ for $\lambda \neq 0$, and extend the definition by continuity to

$$G(0) := \lim_{\lambda \to 0} G(\lambda) = \mathbb{E}[X]. \tag{3.9}$$

Note that we have $G'(\lambda) = \frac{1}{\lambda}\frac{\varphi'(\lambda)}{\varphi(\lambda)} - \frac{1}{\lambda^2}\log\varphi(\lambda)$, so that the inequality (3.8) can be rewritten in the simple form $G'(\lambda) \leq \frac{1}{2}\sigma^2$ for all $\lambda \in I$. For any $\lambda_0 > 0$, we can integrate both sides of the inequality to obtain

$$G(\lambda) - G(\lambda_0) \leq \tfrac{1}{2}\sigma^2(\lambda - \lambda_0).$$

Letting $\lambda_0 \to 0^+$ and using the relation (3.9), we conclude that

$$G(\lambda) - \mathbb{E}[X] \leq \tfrac{1}{2}\sigma^2\lambda,$$

which is equivalent to the claim (3.6). We leave the extension of this proof to the case $I = \mathbb{R}$ as an exercise for the reader. □

Thus far, we have seen how a particular upper bound (3.5) on the entropy $\mathbb{H}(e^{\lambda X})$ translates into a bound on the cumulant generating function (3.6), and hence into sub-Gaussian tail bounds via the usual Chernoff argument. It is natural to explore to what extent this approach may be generalized. As seen previously in Chapter 2, a broader class of random variables are those with sub-exponential tails, and the following result is the analog of Proposition 3.2 in this case.

Proposition 3.3 (Bernstein entropy bound) *Suppose that there are positive constants b and σ such that the entropy $\mathbb{H}(e^{\lambda X})$ satisfies the bound*

$$\mathbb{H}(e^{\lambda X}) \leq \lambda^2\{b\varphi_x'(\lambda) + \varphi_x(\lambda)(\sigma^2 - b\mathbb{E}[X])\} \qquad \text{for all } \lambda \in [0, 1/b). \tag{3.10}$$

Then X satisfies the bound

$$\log \mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \leq \sigma^2\lambda^2(1 - b\lambda)^{-1} \qquad \text{for all } \lambda \in [0, 1/b). \tag{3.11}$$

Remarks: As a consequence of the usual Chernoff argument, Proposition 3.3 implies that X satisfies the upper tail bound

$$\mathbb{P}[X \geq \mathbb{E}[X] + \delta] \leq \exp\left(-\frac{\delta^2}{4\sigma^2 + 2b\delta}\right) \qquad \text{for all } \delta \geq 0, \tag{3.12}$$

which (modulo non-optimal constants) is the usual Bernstein-type bound to be expected for a variable with sub-exponential tails. See Proposition 2.10 from Chapter 2 for further details on such Bernstein bounds.

We now turn to the proof of Proposition 3.3.

Proof As before, we omit the dependence of φ_x on X throughout this proof so as to simplify notation. By rescaling and recentering arguments sketched out in Exercise 3.6, we may assume without loss of generality that $\mathbb{E}[X] = 0$ and $b = 1$, in which case the inequality (3.10) simplifies to

$$\mathbb{H}(e^{\lambda X}) \leq \lambda^2\{\varphi'(\lambda) + \varphi(\lambda)\sigma^2\} \qquad \text{for all } \lambda \in [0, 1). \tag{3.13}$$

Recalling the function $G(\lambda) = \frac{1}{\lambda}\log\varphi(\lambda)$ from the proof of Proposition 3.2, a little bit of

https://doi.org/10.1017/9781108627771.003 Published online by Cambridge University Press

algebra shows that condition (3.13) is equivalent to the differential inequality $G' \leq \sigma^2 + \frac{\varphi'}{\varphi}$. Letting $\lambda_0 > 0$ be arbitrary and integrating both sides of this inequality over the interval (λ_0, λ), we obtain

$$G(\lambda) - G(\lambda_0) \leq \sigma^2(\lambda - \lambda_0) + \log \varphi(\lambda) - \log \varphi(\lambda_0).$$

Since this inequality holds for all $\lambda_0 > 0$, we may take the limit as $\lambda_0 \to 0^+$. Doing so and using the facts that $\lim_{\lambda_0 \to 0^+} G(\lambda_0) = G(0) = \mathbb{E}[X]$ and $\log \varphi(0) = 0$, we obtain the bound

$$G(\lambda) - \mathbb{E}[X] \leq \sigma^2 \lambda + \log \varphi(\lambda). \tag{3.14}$$

Substituting the definition of G and rearranging yields the claim (3.11).

□

3.1.3 Separately convex functions and the entropic method

Thus far, we have seen how the entropic method can be used to derive sub-Gaussian and sub-exponential tail bounds for scalar random variables. If this were the only use of the entropic method, then we would have gained little beyond what can be done via the usual Chernoff bound. The real power of the entropic method—as we now will see—manifests itself in dealing with concentration for functions of many random variables.

As an illustration, we begin by stating a deep result that can be proven in a relatively direct manner using the entropy method. We say that a function $f \colon \mathbb{R}^n \to \mathbb{R}$ is *separately convex* if, for each index $k \in \{1, 2, \ldots, n\}$, the univariate function

$$y_k \mapsto f(x_1, x_2, \ldots, x_{k-1}, y_k, x_{k+1}, \ldots, x_n)$$

is convex for each fixed vector $(x_1, x_2, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n) \in \mathbb{R}^{n-1}$. A function f is L-Lipschitz with respect to the Euclidean norm if

$$|f(x) - f(x')| \leq L \, \|x - x'\|_2 \qquad \text{for all } x, x' \in \mathbb{R}^n. \tag{3.15}$$

The following result applies to separately convex and L-Lipschitz functions.

Theorem 3.4 *Let $\{X_i\}_{i=1}^n$ be independent random variables, each supported on the interval $[a, b]$, and let $f \colon \mathbb{R}^n \to \mathbb{R}$ be separately convex, and L-Lipschitz with respect to the Euclidean norm. Then, for all $\delta > 0$, we have*

$$\mathbb{P}[f(X) \geq \mathbb{E}[f(X)] + \delta] \leq \exp\left(-\frac{\delta^2}{4L^2(b-a)^2}\right). \tag{3.16}$$

Remarks: This result is the analog of the upper tail bound for Lipschitz functions of Gaussian variables (cf. Theorem 2.26 in Chapter 2), but applicable to independent and bounded variables instead. In contrast to the Gaussian case, the additional assumption of separate convexity cannot be eliminated in general; see the bibliographic section for further discussion. When f is jointly convex, other techniques can be used to obtain the lower tail bound as well; see Theorem 3.24 in the sequel for one such example.

Theorem 3.4 can be used to obtain order-optimal bounds for a number of interesting problems. As one illustration, we return to the Rademacher complexity, first introduced in Example 2.25 of Chapter 2.

Example 3.5 (Sharp bounds on Rademacher complexity) Given a bounded subset $\mathcal{A} \subset \mathbb{R}^n$, consider the random variable $Z = \sup_{a\in\mathcal{A}} \sum_{k=1}^n a_k \varepsilon_k$, where $\varepsilon_k \in \{-1, +1\}$ are i.i.d. Rademacher variables. Let us view Z as a function of the random signs, and use Theorem 3.4 to bound the probability of the tail event $\{Z \geq \mathbb{E}[Z] + t\}$.

It suffices to verify the convexity and Lipschitz conditions of the theorem. First, since $Z = Z(\varepsilon_1, \ldots, \varepsilon_n)$ is the maximum of a collection of linear functions, it is jointly (and hence separately) convex. Let $Z' = Z(\varepsilon'_1, \ldots, \varepsilon'_n)$ where $\varepsilon' \in \{-1, +1\}^n$ is a second vector of sign variables. For any $a \in \mathcal{A}$, we have

$$\underbrace{\langle a,\, \varepsilon\rangle}_{\sum_{k=1}^n a_k\varepsilon_k} - Z' = \langle a,\, \varepsilon\rangle - \sup_{a'\in\mathcal{A}} \langle a',\, \varepsilon'\rangle \leq \langle a,\, \varepsilon - \varepsilon'\rangle \leq \|a\|_2\, \|\varepsilon - \varepsilon'\|_2.$$

Taking suprema over $a \in \mathcal{A}$ yields that $Z - Z' \leq (\sup_{a\in\mathcal{A}} \|a\|_2)\, \|\varepsilon - \varepsilon'\|_2$. Since the same argument may be applied with the roles of ε and ε' reversed, we conclude that Z is Lipschitz with parameter $\mathcal{W}(\mathcal{A}) := \sup_{a\in\mathcal{A}} \|a\|_2$, corresponding to the Euclidean width of the set. Putting together the pieces, Theorem 3.4 implies that

$$\mathbb{P}[Z \geq \mathbb{E}[Z] + t] \leq \exp\left(-\frac{t^2}{16\mathcal{W}^2(\mathcal{A})}\right). \tag{3.17}$$

Note that parameter $\mathcal{W}^2(\mathcal{A})$ may be substantially smaller than the quantity $\sum_{k=1}^n \sup_{a\in\mathcal{A}} a_k^2$ —indeed, possibly as much as a factor of n smaller! In such cases, Theorem 3.4 yields a much sharper tail bound than our earlier tail bound from Example 2.25, which was obtained by applying the bounded differences inequality. ♣

Another use of Theorem 3.4 is in random matrix theory.

Example 3.6 (Operator norm of a random matrix) Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be a random matrix, say with X_{ij} drawn i.i.d. from some zero-mean distribution supported on the unit interval $[-1, +1]$. The spectral or ℓ_2-operator norm of X, denoted by $|||\mathbf{X}|||_2$, is its maximum singular value, given by

$$|||\mathbf{X}|||_2 = \max_{\substack{v\in\mathbb{R}^d \\ \|v\|_2=1}} \|\mathbf{X}v\|_2 = \max_{\substack{v\in\mathbb{R}^d \\ \|v\|_2=1}} \max_{\substack{u\in\mathbb{R}^n \\ \|u\|_2=1}} u^{\mathrm{T}}\mathbf{X}v. \tag{3.18}$$

Let us view the mapping $\mathbf{X} \mapsto |||\mathbf{X}|||_2$ as a function f from $\mathbb{R}^{nd}$ to $\mathbb{R}$. In order to apply Theorem 3.4, we need to show that f is both Lipschitz and convex. From its definition (3.18), the operator norm is the supremum of a collection of functions that are linear in the entries $\mathbf{X}$; any such supremum is a convex function. Moreover, we have

$$\Big| |||\mathbf{X}|||_2 - |||\mathbf{X}'|||_2 \Big| \overset{\text{(i)}}{\leq} |||\mathbf{X} - \mathbf{X}'|||_2 \overset{\text{(ii)}}{\leq} |||\mathbf{X} - \mathbf{X}'|||_{\mathrm{F}}, \tag{3.19}$$

where step (i) follows from the triangle inequality, and step (ii) follows since the Frobenius

https://doi.org/10.1017/9781108627771.003 Published online by Cambridge University Press

norm of a matrix always upper bounds the operator norm. (The Frobenius norm $|||\mathbf{M}|||_F$ of a matrix $\mathbf{M} \in \mathbb{R}^{n\times d}$ is simply the Euclidean norm of all its entries; see equation (2.50).) Consequently, the operator norm is Lipschitz with parameter $L = 1$, and thus Theorem 3.4 implies that

$$\mathbb{P}[|||\mathbf{X}|||_2 \geq \mathbb{E}[|||\mathbf{X}|||_2] + \delta] \leq e^{-\frac{\delta^2}{16}}.$$

It is worth observing that this bound is the analog of our earlier bound (2.52) on the operator norm of a Gaussian random matrix, albeit with a worse constant. See Example 2.32 in Chapter 2 for further details on this Gaussian case. ♣

3.1.4 Tensorization and separately convex functions

We now return to prove Theorem 3.4. The proof is based on two lemmas, both of which are of independent interest. Here we state these results and discuss some of their consequences, deferring their proofs to the end of this section. Our first lemma establishes an entropy bound for univariate functions:

Lemma 3.7 (Entropy bound for univariate functions) *Let $X, Y \sim \mathbb{P}$ be a pair of i.i.d. variates. Then for any function $g\colon \mathbb{R} \to \mathbb{R}$, we have*

$$\mathbb{H}(e^{\lambda g(X)}) \leq \lambda^2 \mathbb{E}[(g(X) - g(Y))^2 e^{\lambda g(X)}\, \mathbb{I}[g(X) \geq g(Y)]] \qquad \textit{for all } \lambda > 0. \tag{3.20a}$$

If in addition X is supported on $[a, b]$, and g is convex and Lipschitz, then

$$\mathbb{H}(e^{\lambda g(X)}) \leq \lambda^2 (b-a)^2\, \mathbb{E}[(g'(X))^2 e^{\lambda g(X)}] \qquad \textit{for all } \lambda > 0, \tag{3.20b}$$

where g' is the derivative.

In stating this lemma, we have used the fact that any convex and Lipschitz function has a derivative defined almost everywhere, a result known as Rademacher's theorem. Moreover, note that if g is Lipschitz with parameter L, then we are guaranteed that $\|g'\|_\infty \leq L$, so that inequality (3.20b) implies an entropy bound of the form

$$\mathbb{H}(e^{\lambda g(X)}) \leq \lambda^2 L^2 (b-a)^2\, \mathbb{E}[e^{\lambda g(X)}] \qquad \text{for all } \lambda > 0.$$

In turn, by an application of Proposition 3.2, such an entropy inequality implies the upper tail bound

$$\mathbb{P}[g(X) \geq \mathbb{E}[g(X)] + \delta] \leq e^{-\frac{\delta^2}{4L^2(b-a)^2}}.$$

Thus, Lemma 3.7 implies the univariate version of Theorem 3.4. However, the inequality (3.20b) is sharper, in that it involves $g'(X)$ as opposed to the worst-case bound L, and this distinction will be important in deriving the sharp result of Theorem 3.4. The more general inequality (3.20b) will be useful in deriving functional versions of the Hoeffding and Bernstein inequalities (see Section 3.4).

https://doi.org/10.1017/9781108627771.003 Published online by Cambridge University Press

Returning to the main thread, it remains to extend this univariate result to the multivariate setting, and the so-called *tensorization property* of entropy plays a key role here. Given a function $f\colon \mathbb{R}^n \to \mathbb{R}$, an index $k \in \{1, 2, \ldots, n\}$ and a vector $x_{\backslash k} = (x_i, i \neq k) \in \mathbb{R}^{n-1}$, we define the conditional entropy in coordinate k via

$$\mathbb{H}(e^{\lambda f_k(X_k)} \mid x_{\backslash k}) := \mathbb{H}(e^{\lambda f(x_1,\ldots,x_{k-1},X_k,x_{k+1},\ldots,x_n)}),$$

where $f_k\colon \mathbb{R} \to \mathbb{R}$ is the coordinate function $x_k \mapsto f(x_1, \ldots, x_k, \ldots, x_n)$. To be clear, for a random vector $X^{\backslash k} \in \mathbb{R}^{n-1}$, the entropy $\mathbb{H}(e^{\lambda f_k(X_k)} \mid X^{\backslash k})$ is a random variable, and its expectation is often referred to as the conditional entropy.) The following result shows that the joint entropy can be upper bounded by a sum of univariate entropies, suitably defined.

Lemma 3.8 (Tensorization of entropy) *Let $f\colon \mathbb{R}^n \to \mathbb{R}$, and let $\{X_k\}_{k=1}^n$ be independent random variables. Then*

$$\mathbb{H}(e^{\lambda f(X_1,\ldots,X_n)}) \leq \mathbb{E}\left[\sum_{k=1}^n \mathbb{H}(e^{\lambda f_k(X_k)} \mid X^{\backslash k})\right] \qquad \text{for all } \lambda > 0. \tag{3.21}$$

Equipped with these two results, we are now ready to prove Theorem 3.4.

Proof of Theorem 3.4 For any $k \in \{1, 2, \ldots, n\}$ and fixed vector $x_{\backslash k} \in \mathbb{R}^{n-1}$, our assumptions imply that the coordinate function f_k is convex, and hence Lemma 3.7 implies that, for all $\lambda > 0$, we have

$$\begin{aligned}\mathbb{H}(e^{\lambda f_k(X_k)} \mid x_{\backslash k}) &\leq \lambda^2 (b-a)^2\, \mathbb{E}_{X_k}[(f_k'(X_k))^2 e^{\lambda f_k(X_k)} \mid x_{\backslash k}] \\ &= \lambda^2 (b-a)^2\, \mathbb{E}_{X_k}\left[\left(\frac{\partial f(x_1, \ldots, X_k, \ldots, x_n)}{\partial x_k}\right)^2 e^{\lambda f(x_1,\ldots,X_k,\ldots,x_n)}\right],\end{aligned}$$

where the second line involves unpacking the definition of the conditional entropy.

Combined with Lemma 3.8, we find that the unconditional entropy is upper bounded as

$$\mathbb{H}(e^{\lambda f(X)}) \leq \lambda^2 (b-a)^2\, \mathbb{E}\left[\sum_{k=1}^n \left(\frac{\partial f(X)}{\partial x_k}\right)^2 e^{\lambda f(X)}\right] \overset{\text{(i)}}{\leq} \lambda^2 (b-a)^2 L^2\, \mathbb{E}[e^{\lambda f(X)}].$$

Here step (i) follows from the Lipschitz condition, which guarantees that

$$\|\nabla f(x)\|_2^2 = \sum_{k=1}^n \left(\frac{\partial f(x)}{\partial x_k}\right)^2 \leq L^2$$

almost surely. Thus, the tail bound (3.16) follows from an application of Proposition 3.2. □

It remains to prove the two auxiliary lemmas used in the preceding proof—namely, Lemma 3.7 on entropy bounds for univariate Lipschitz functions, and Lemma 3.8 on the tensorization of entropy. We begin with the former property.

https://doi.org/10.1017/9781108627771.003 Published online by Cambridge University Press

Proof of Lemma 3.7

By the definition of entropy, we can write

$$\begin{aligned}\mathbb{H}(e^{\lambda g(X)}) &= \mathbb{E}_X[\lambda g(X)e^{\lambda g(X)}] - \mathbb{E}_X[e^{\lambda g(X)}]\log\left(\mathbb{E}_Y[e^{\lambda g(Y)}]\right)\\ &\overset{\text{(i)}}{\leq} \mathbb{E}_X[\lambda g(X)e^{\lambda g(X)}] - \mathbb{E}_{X,Y}[e^{\lambda g(X)}\lambda g(Y)]\\ &= \tfrac{1}{2}\mathbb{E}_{X,Y}\Big[\lambda\left\{g(X)-g(Y)\right\}\left\{e^{\lambda g(X)} - e^{\lambda g(Y)}\right\}\Big]\\ &\overset{\text{(ii)}}{=} \lambda\mathbb{E}\Big[\{g(X)-g(Y)\}\{e^{\lambda g(X)} - e^{\lambda g(Y)}\}\,\mathbb{I}[g(X)\geq g(Y)]\Big],\end{aligned} \tag{3.22}$$

where step (i) follows from Jensen's inequality, and step (ii) follows from symmetry of X and Y.

By convexity of the exponential, we have $e^s - e^t \leq e^s(s-t)$ for all $s, t \in \mathbb{R}$. For $s \geq t$, we can multiply both sides by $(s-t) \geq 0$, thereby obtaining

$$(s-t)(e^s - e^t)\,\mathbb{I}[s \geq t] \leq (s-t)^2 e^s\,\mathbb{I}[s \geq t].$$

Applying this bound with $s = \lambda g(X)$ and $t = \lambda g(Y)$ to the inequality (3.22) yields

$$\mathbb{H}(e^{\lambda g(X)}) \leq \lambda^2\,\mathbb{E}[(g(X)-g(Y))^2 e^{\lambda g(X)}\,\mathbb{I}[g(X)\geq g(Y)]], \tag{3.23}$$

where we have recalled the assumption that $\lambda > 0$.

If in addition g is convex, then we have the upper bound $g(x) - g(y) \leq g'(x)(x-y)$, and hence, for $g(x) \geq g(y)$,

$$(g(x)-g(y))^2 \leq (g'(x))^2(x-y)^2 \leq (g'(x))^2(b-a)^2,$$

where the final step uses the assumption that $x, y \in [a, b]$. Combining the pieces yields the claim.

We now turn to the tensorization property of entropy.

Proof of Lemma 3.8

The proof makes use of the following variational representation for entropy:

$$\mathbb{H}(e^{\lambda f(X)}) = \sup_g\{\mathbb{E}[g(X)e^{\lambda f(X)}] \mid \mathbb{E}[e^{g(X)}] \leq 1\}. \tag{3.24}$$

This equivalence follows by a duality argument that we explore in Exercise 3.9.

For each $j \in \{1, 2, \ldots, n\}$, define $X_j^n = (X_j, \ldots, X_n)$. Let g be any function that satisfies $\mathbb{E}[e^{g(X)}] \leq 1$. We can then define an auxiliary sequence of functions $\{g^1, \ldots, g^n\}$ via

$$g^1(X_1, \ldots, X_n) := g(X) - \log\mathbb{E}[e^{g(X)} \mid X_2^n]$$

and

$$g^k(X_k, \ldots, X_n) := \log\frac{\mathbb{E}[e^{g(X)} \mid X_k^n]}{\mathbb{E}[e^{g(X)} \mid X_{k+1}^n]} \qquad \text{for } k = 2, \ldots, n.$$

By construction, we have

$$\sum_{k=1}^n g^k(X_k, \ldots, X_n) = g(X) - \log\mathbb{E}[e^{g(X)}] \geq g(X) \tag{3.25}$$

and moreover $\mathbb{E}[\exp(g^k(X_k, X_{k+1}, \ldots, X_n)) \mid X_{k+1}^n] = 1$.

We now use this decomposition within the variational representation (3.24), thereby obtaining the chain of upper bounds

$$\begin{aligned}\mathbb{E}[g(X)e^{\lambda f(X)}] &\overset{\text{(i)}}{\leq} \sum_{k=1}^{n} \mathbb{E}[g^k(X_k, \ldots, X_n)e^{\lambda f(X)}] \\ &= \sum_{k=1}^{n} \mathbb{E}_{X_{\setminus k}}[\mathbb{E}_{X_k}[g^k(X_k, \ldots, X_n)e^{\lambda f(X)} \mid X_{\setminus k}]] \\ &\overset{\text{(ii)}}{\leq} \sum_{k=1}^{n} \mathbb{E}_{X_{\setminus k}}[\mathbb{H}(e^{\lambda f_k(X_k)} \mid X_{\setminus k})],\end{aligned}$$

where inequality (i) uses the bound (3.25), and inequality (ii) applies the variational representation (3.24) to the univariate functions, and also makes use of the fact that $\mathbb{E}[g^k(X_k, \ldots, X_n) \mid X_{\setminus k}] = 1$. Since this argument applies to any function g such that $\mathbb{E}[e^{g(X)}] \leq 1$, we may take the supremum over the left-hand side, and combined with the variational representation (3.24), we conclude that

$$\mathbb{H}(e^{\lambda f(X)}) \leq \sum_{k=1}^{n} \mathbb{E}_{X_{\setminus k}}[\mathbb{H}(e^{\lambda f_k(X_k)} \mid X_{\setminus k})],$$

as claimed.

3.2 A geometric perspective on concentration

We now turn to some geometric aspects of the concentration of measure. Historically, this geometric viewpoint is among the oldest, dating back to the classical result of Lévy on concentration of measure for Lipschitz functions of Gaussians. It also establishes deep links between probabilistic concepts and high-dimensional geometry.

The results of this section are most conveniently stated in terms of a *metric measure space*—namely, a metric space $(\mathcal{X}, \rho)$ endowed with a probability measure $\mathbb{P}$ on its Borel sets. Some canonical examples of metric spaces for the reader to keep in mind are the set $\mathcal{X} = \mathbb{R}^n$ equipped with the usual Euclidean metric $\rho(x, y) := \|x - y\|_2$, and the discrete cube $\mathcal{X} = \{0, 1\}^n$ equipped with the Hamming metric $\rho(x, y) = \sum_{j=1}^{n} \mathbb{I}[x_j \neq y_j]$.

Associated with any metric measure space is an object known as its *concentration function*, which is defined in a geometric manner via the ϵ-enlargements of sets. The concentration function specifies how rapidly, as a function of ϵ, the probability of any ϵ-enlargement increases towards one. As we will see, this function is intimately related to the concentration properties of Lipschitz functions on the metric space.

3.2.1 Concentration functions

Given a set $A \subseteq \mathcal{X}$ and a point $x \in \mathcal{X}$, define the quantity

$$\rho(x, A) := \inf_{y \in A} \rho(x, y), \tag{3.26}$$

https://doi.org/10.1017/9781108627771.003 Published online by Cambridge University Press

which measures the distance between the point x and the closest point in the set A. Given a parameter $\epsilon > 0$, the *ϵ-enlargement* of A is given by

$$A^{\epsilon} := \{x \in \mathcal{X} \mid \rho(x, A) < \epsilon\}. \tag{3.27}$$

In words, the set A^{ϵ} corresponds to the open neighborhood of points lying at distance less than ϵ from A. With this notation, the concentration function of the metric measure space $(\mathcal{X}, \rho, \mathbb{P})$ is defined as follows:

Definition 3.9 The *concentration function* $\alpha \colon [0, \infty) \to \mathbb{R}_{+}$ associated with metric measure space $(\mathbb{P}, \mathcal{X}, \rho)$ is given by

$$\alpha_{\mathbb{P},(\mathcal{X},\rho)}(\epsilon) := \sup_{A \subseteq \mathcal{X}} \{1 - \mathbb{P}[A^{\epsilon}] \mid \mathbb{P}[A] \geq \tfrac{1}{2}\}, \tag{3.28}$$

where the supremum is taken over all measurable subsets A.

When the underlying metric space $(\mathcal{X}, \rho)$ is clear from the context, we frequently use the abbreviated notation $\alpha_{\mathbb{P}}$. It follows immediately from the definition (3.28) that $\alpha_{\mathbb{P}}(\epsilon) \in [0, \frac{1}{2}]$ for all $\epsilon \geq 0$. Of primary interest is the behavior of the concentration function as ϵ increases, and, more precisely, how rapidly it approaches zero. Let us consider some examples to illustrate.

Example 3.10 (Concentration function for sphere) Consider the metric measure space defined by the uniform distribution over the n-dimensional Euclidean sphere

$$\mathbb{S}^{n-1} := \{x \in \mathbb{R}^{n} \mid \|x\|_2 = 1\}, \tag{3.29}$$

equipped with the geodesic distance $\rho(x, y) := \arccos \langle x,\ y\rangle$. Let us upper bound the concentration function $\alpha_{\mathbb{S}^{n-1}}$ defined by the triplet $(\mathbb{P}, \mathbb{S}^{n-1}, \rho)$, where $\mathbb{P}$ is the uniform distribution over the sphere. For each $y \in \mathbb{S}^{n-1}$, we can define the hemisphere

$$H_y := \{x \in \mathbb{S}^{n-1} \mid \rho(x, y) \geq \pi/2\} = \{x \in \mathbb{S}^{n-1} \mid \langle x,\ y\rangle \leq 0\}, \tag{3.30}$$

as illustrated in Figure 3.1(a). With some simple geometry, it can be shown that its ϵ-enlargement corresponds to the set

$$H_y^{\epsilon} = \{z \in \mathbb{S}^{n-1} \mid \langle z,\ y\rangle < \sin(\epsilon)\}, \tag{3.31}$$

as illustrated in Figure 3.1(b). Note that $\mathbb{P}[H_y] = 1/2$, so that the hemisphere (3.30) is a candidate set for the supremum defining the concentration function (3.28). The classical isoperimetric theorem of Lévy asserts that these hemispheres are *extremal*, meaning that they achieve the supremum, viz.

$$\alpha_{\mathbb{S}^{n-1}}(\epsilon) = 1 - \mathbb{P}[H_y^{\epsilon}]. \tag{3.32}$$

Let us take this fact as given, and use it to compute an upper bound on the concentration

https://doi.org/10.1017/9781108627771.003 Published online by Cambridge University Press

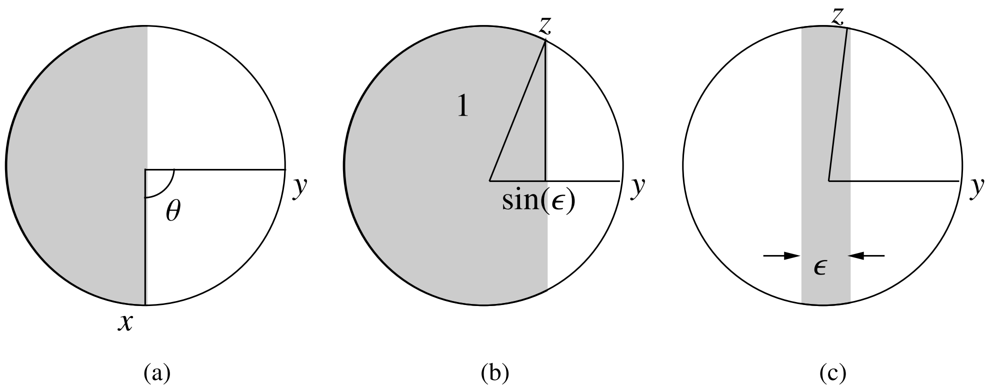

Figure 3.1 (a) Idealized illustration of the sphere $\mathbb{S}^{n-1}$. Any vector $y \in \mathbb{S}^{n-1}$ defines a hemisphere $H_y = \{x \in \mathbb{S}^{n-1} \mid \langle x, y\rangle \leq 0\}$, corresponding to those vectors whose angle $\theta = \arccos\langle x, y\rangle$ with y is at least $\pi/2$ radians. (b) The ϵ-enlargement of the hemisphere H_y. (c) A central slice $T_y(\epsilon)$ of the sphere of width ϵ.

function. In order to do so, we need to lower bound the probability $\mathbb{P}[H_y^\epsilon]$. Since $\sin(\epsilon) \geq \epsilon/2$ for all $\epsilon \in (0, \pi/2]$, the enlargement contains the set

$$\widetilde{H}_y^\epsilon := \{z \in \mathbb{S}^{n-1} \mid \langle z, y\rangle \leq \tfrac{1}{2}\epsilon\},$$

and hence $\mathbb{P}[H_y^\epsilon] \geq \mathbb{P}[\widetilde{H}_y^\epsilon]$. Finally, a geometric calculation, left as an exercise for the reader, yields that, for all $\epsilon \in (0, \sqrt{2})$, we have

$$\mathbb{P}[\widetilde{H}_y^\epsilon] \geq 1 - \left(1 - \left(\frac{\epsilon}{2}\right)^2\right)^{n/2} \geq 1 - e^{-n\epsilon^2/8}, \tag{3.33}$$

where we have used the inequality $(1 - t) \leq e^{-t}$ with $t = \epsilon^2/4$. We thus obtain that the concentration function is upper bounded as $\alpha_{\mathbb{S}^{n-1}}(\epsilon) \leq e^{-n\epsilon^2/8}$. A similar but more careful approach to bounding $\mathbb{P}[H_y]$ can be used to establish the sharper upper bound

$$\alpha_{\mathbb{S}^{n-1}}(\epsilon) \leq \sqrt{\frac{\pi}{2}}\, e^{-\frac{n\epsilon^2}{2}}. \tag{3.34}$$

The bound (3.34) is an extraordinary conclusion, originally due to Lévy, and it is worth pausing to think about it in more depth. Among other consequences, it implies that, if we consider a central slice of the sphere of width ϵ, say a set of the form

$$T_y(\epsilon) := \{z \in \mathbb{S}^{n-1} \mid |\langle z, y\rangle| \leq \epsilon/2\}, \tag{3.35}$$

as illustrated in Figure 3.1(c), then it occupies a huge fraction of the total volume: in particular, we have $\mathbb{P}[T_y(\epsilon)] \geq 1 - \sqrt{2\pi}\exp(-\frac{n\epsilon^2}{2})$. Moreover, this conclusion holds for *any* such slice. To be clear, the two-dimensional instance shown in Figure 3.1(c)—like any low-dimensional example—fails to capture the behavior of high-dimensional spheres. In general, our low-dimensional intuition can be *very* misleading when applied to high-dimensional settings. ♣

https://doi.org/10.1017/9781108627771.003 Published online by Cambridge University Press

3.2.2 Connection to Lipschitz functions

In Chapter 2 and the preceding section of this chapter, we explored some methods for obtaining deviation and concentration inequalities for various types of Lipschitz functions. The concentration function $\alpha_{\mathbb{P},(\mathcal{X},\rho)}$ turns out to be intimately related to such results on the tail behavior of Lipschitz functions. In particular, suppose that a function $f\colon \mathcal{X} \to \mathbb{R}$ is L-Lipschitz with respect to the metric ρ—that is,

$$|f(x) - f(y)| \leq L\rho(x, y) \qquad \text{for all } x, y \in \mathcal{X}. \tag{3.36}$$

Given a random variable $X \sim \mathbb{P}$, let m_f be any median of $f(X)$, meaning a number such that

$$\mathbb{P}[f(X) \geq m_f] \geq 1/2 \quad \text{and} \quad \mathbb{P}[f(X) \leq m_f] \geq 1/2. \tag{3.37}$$

Define the set $A = \{x \in \mathcal{X} \mid f(x) \leq m_f\}$, and consider its $\frac{\epsilon}{L}$-enlargement $A^{\epsilon/L}$. For any $x \in A^{\epsilon/L}$, there exists some $y \in A$ such that $\rho(x, y) < \epsilon/L$. Combined with the Lipschitz property, we conclude that $|f(y) - f(x)| \leq L\rho(x, y) < \epsilon$, and hence that

$$A^{\epsilon/L} \subseteq \{x \in \mathcal{X} \mid f(x) < m_f + \epsilon\}. \tag{3.38}$$

Consequently, we have

$$\mathbb{P}[f(X) \geq m_f + \epsilon] \overset{\text{(i)}}{\leq} 1 - \mathbb{P}[A^{\epsilon/L}] \overset{\text{(ii)}}{\leq} \alpha_{\mathbb{P}}(\epsilon/L),$$

where inequality (i) follows from the inclusion (3.38), and inequality (ii) uses the fact $\mathbb{P}[A] \geq 1/2$, and the definition (3.28). Applying a similar argument to $-f$ yields an analogous left-sided deviation inequality $\mathbb{P}[f(X) \leq m_f - \epsilon] \leq \alpha_{\mathbb{P}}(\epsilon/L)$, and putting together the pieces yields the concentration inequality

$$\mathbb{P}[|f(X) - m_f| \geq \epsilon] \leq 2\alpha_{\mathbb{P}}(\epsilon/L).$$

As shown in Exercise 2.14 from Chapter 2, such sharp concentration around the median is equivalent (up to constant factors) to concentration around the mean. Consequently, we have shown that bounds on the concentration function (3.28) imply concentration inequalities for any Lipschitz function. This argument can also be reversed, yielding the following equivalence between control on the concentration function, and the behavior of Lipschitz functions.

Proposition 3.11 *Given a random variable $X \sim \mathbb{P}$ and concentration function $\alpha_{\mathbb{P}}$, any 1-Lipschitz function on $(\mathcal{X}, \rho)$ satisfies*

$$\mathbb{P}[|f(X) - m_f| \geq \epsilon] \leq 2\alpha_{\mathbb{P}}(\epsilon), \tag{3.39a}$$

where m_f is any median of f. Conversely, suppose that there is a function $\beta\colon \mathbb{R}_+ \to \mathbb{R}_+$ such that, for any 1-Lipschitz function on $(\mathcal{X}, \rho)$,

$$\mathbb{P}[f(X) \geq \mathbb{E}[f(X)] + \epsilon] \leq \beta(\epsilon) \qquad \textit{for all } \epsilon \geq 0. \tag{3.39b}$$

Then the concentration function satisfies the bound $\alpha_{\mathbb{P}}(\epsilon) \leq \beta(\epsilon/2)$.

Proof It remains to prove the converse claim. Fix some $\epsilon \geq 0$, and let A be an arbitrary measurable set with $\mathbb{P}[A] \geq 1/2$. Recalling the definition of $\rho(x, A)$ from equation (3.26), let us consider the function $f(x) := \min\{\rho(x, A), \epsilon\}$. It can be seen that f is 1-Lipschitz, and moreover that $1 - \mathbb{P}[A^\epsilon] = \mathbb{P}[f(X) \geq \epsilon]$. On the other hand, our construction guarantees that

$$\mathbb{E}[f(X)] \leq (1 - \mathbb{P}[A])\epsilon \leq \epsilon/2,$$

whence we have

$$\mathbb{P}[f(X) \geq \epsilon] \leq \mathbb{P}[f(X) \geq \mathbb{E}[f(X)] + \epsilon/2] \leq \beta(\epsilon/2),$$

where the final inequality uses the assumed condition (3.39b). □

Proposition 3.11 has a number of concrete interpretations in specific settings.

Example 3.12 (Lévy concentration on $\mathbb{S}^{n-1}$) From our earlier discussion in Example 3.10, the concentration function for the uniform distribution over the sphere $\mathbb{S}^{n-1}$ can be upper bounded as

$$\alpha_{\mathbb{S}^{n-1}}(\epsilon) \leq \sqrt{\frac{\pi}{2}}\, e^{-\frac{n\epsilon^2}{2}}.$$

Consequently, for any 1-Lipschitz function f defined on the sphere $\mathbb{S}^{n-1}$, we have the two-sided bound

$$\mathbb{P}[|f(X) - m_f| \geq \epsilon] \leq \sqrt{2\pi}\, e^{-\frac{n\epsilon^2}{2}}, \tag{3.40}$$

where m_f is any median of f. Moreover, by the result of Exercise 2.14(d), we also have

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq \epsilon] \leq 2\sqrt{2\pi}\, e^{-\frac{n\epsilon^2}{8}}. \tag{3.41}$$

♣

Example 3.13 (Concentration for Boolean hypercube) Consider the Boolean hypercube $\mathcal{X} = \{0, 1\}^n$ equipped with the usual Hamming metric

$$\rho_H(x, y) := \sum_{j=1}^{n} \mathbb{I}[x_j \neq y_j].$$

Given this metric, we can define the Hamming ball

$$\mathbb{B}_H(r; x) = \{y \in \{0, 1\}^n \mid \rho_H(y, x) \leq r\}$$

of radius r centered at some $x \in \{0, 1\}^n$. Of interest here are the Hamming balls centered at the all-zeros vector 0 and all-ones vector 1, respectively. In particular, in this example, we show how a classical combinatorial result due to Harper can be used to bound the concentration function of the metric measure space consisting of the Hamming metric along with the uniform distribution $\mathbb{P}$.

Given two non-empty subsets A and B of the binary hypercube, one consequence of Harper's theorem is that we can always find two positive integers r_A and r_B, and associated subsets A' and B', with the following properties:

- the sets A' and B' are sandwiched as

$$\mathbb{B}_H(r_A - 1; 0) \subseteq A' \subseteq \mathbb{B}_H(r_A; 0) \quad \text{and} \quad \mathbb{B}_H(r_B - 1; 1) \subseteq B' \subseteq \mathbb{B}_H(r_B; 1);$$

https://doi.org/10.1017/9781108627771.003 Published online by Cambridge University Press

- the cardinalities are matched as $\text{card}(A) = \text{card}(A')$ and $\text{card}(B) = \text{card}(B')$;
- we have the lower bound $\rho_H(A', B') \geq \rho_H(A, B)$.

Let us now show that this combinatorial theorem implies that

$$\alpha_{\mathbb{P}}(\epsilon) \leq e^{-\frac{2\epsilon^2}{n}} \qquad \text{for all } n \geq 3. \tag{3.42}$$

Consider any subset such that $\mathbb{P}[A] = \frac{\text{card}(A)}{2^n} \geq \frac{1}{2}$. For any $\epsilon > 0$, define the set $B = \{0,1\}^n \setminus A^\epsilon$. In order to prove the bound (3.42), it suffices to show that $\mathbb{P}[B] \leq e^{-\frac{2\epsilon^2}{n}}$. Since we always have $\mathbb{P}[B] \leq \frac{1}{2} \leq e^{-\frac{2}{n}}$ for $n \geq 3$, it suffices to restrict our attention to $\epsilon > 1$. By construction, we have

$$\rho_H(A, B) = \min_{a \in A, b \in B} \rho_H(a, b) \geq \epsilon.$$

Let A' and B' denote the subsets guaranteed by Harper's theorem. Since A has cardinality at least 2^{n-1}, the set A', which has the same cardinality as A, must contain all vectors with at most $n/2$ ones. Moreover, by the cardinality matching condition and our choice of the uniform distribution, we have $\mathbb{P}[B] = \mathbb{P}[B']$. On the other hand, the set B' is contained within a Hamming ball centered at the all-ones vector, and we have $\rho_H(A', B') \geq \epsilon > 1$. Consequently, any vector $b \in B'$ must contain at least $\frac{n}{2} + \epsilon$ ones. Thus, if we let $\{X_i\}_{i=1}^n$ be a sequence of i.i.d. Bernoulli variables, we have $\mathbb{P}[B'] \leq \mathbb{P}[\sum_{i=1}^n X_i \geq \frac{n}{2} + \epsilon] \leq e^{-\frac{2\epsilon^2}{n}}$, where the final inequality follows from the Hoeffding bound.

Since A was an arbitrary set with $\mathbb{P}[A] \geq \frac{1}{2}$, we have shown that the concentration function satisfies the bound (3.42). Applying Proposition 3.11, we conclude that any 1-Lipschitz function on the Boolean hypercube satisfies the concentration bound

$$\mathbb{P}[|f(X) - m_f| \geq \epsilon] \leq 2e^{-\frac{2\epsilon^2}{n}}.$$

Thus, modulo the negligible difference between the mean and median (see Exercise 2.14), we have recovered the bounded differences inequality (2.35) for Lipschitz functions on the Boolean hypercube. ♣

3.2.3 From geometry to concentration

The geometric perspective suggests the possibility of a variety of connections between convex geometry and the concentration of measure. Consider, for instance, the Brunn–Minkowski inequality: in one of its formulations, it asserts that, for any two convex bodies[2] C and D in $\mathbb{R}^n$, we have

$$[\text{vol}(\lambda C + (1-\lambda)D)]^{1/n} \geq \lambda[\text{vol}(C)]^{1/n} + (1-\lambda)[\text{vol}(D)]^{1/n} \qquad \text{for all } \lambda \in [0,1]. \tag{3.43}$$

Here we use

$$\lambda C + (1-\lambda)D := \{\lambda c + (1-\lambda)d \mid c \in C, d \in D\}$$

to denote the Minkowski sum of the two sets. The Brunn–Minkowski inequality and its variants are intimately connected to concentration of measure. To appreciate the connection,

[2] A convex body in $\mathbb{R}^n$ is a compact and closed set.

https://doi.org/10.1017/9781108627771.003 Published online by Cambridge University Press

observe that the concentration function (3.28) defines a notion of extremal sets—namely, those that minimize the measure $\mathbb{P}[A^\epsilon]$ subject to a constraint on the size of $\mathbb{P}[A]$. Viewing the volume as a type of unnormalized probability measure, the Brunn–Minkowski inequality (3.43) can be used to prove a classical result of this type:

Example 3.14 (Classical isoperimetric inequality in $\mathbb{R}^n$) Consider the Euclidean sphere $\mathbb{B}_2^n := \{x \in \mathbb{R}^n \mid \|x\|_2 \leq 1\}$ in $\mathbb{R}^n$. The classical isoperimetric inequality asserts that, for any set $A \subset \mathbb{R}^n$ such that $\text{vol}(A) = \text{vol}(\mathbb{B}_2^n)$, the volume of its ϵ-enlargement A^ϵ is lower bounded as

$$\text{vol}(A^\epsilon) \geq \text{vol}([\mathbb{B}_2^n]^\epsilon), \tag{3.44}$$

showing that the ball $\mathbb{B}_2^n$ is extremal. In order to verify this bound, we note that

$$[\text{vol}(A^\epsilon)]^{1/n} = [\text{vol}(A + \epsilon\mathbb{B}_2^n)]^{1/n} \geq [\text{vol}(A)]^{1/n} + [\text{vol}(\epsilon\mathbb{B}_2^n)]^{1/n},$$

where the lower bound follows by applying the Brunn–Minkowski inequality (3.43) with appropriate choices of (λ, C, D); see Exercise 3.10 for the details. Since $\text{vol}(A) = \text{vol}(\mathbb{B}_2^n)$ and $[\text{vol}(\epsilon\mathbb{B}_2^n)]^{1/n} = \epsilon\, \text{vol}(\mathbb{B}_2^n)$, we see that

$$\text{vol}(A^\epsilon)^{1/n} \geq (1 + \epsilon)\, \text{vol}(\mathbb{B}_2^n)^{1/n} = [\text{vol}((\mathbb{B}_2^n)^\epsilon)]^{1/n},$$

which establishes the claim. ♣

The Brunn–Minkowski inequality has various equivalent formulations. For instance, it can also be stated as

$$\text{vol}(\lambda C + (1-\lambda)D) \geq [\text{vol}(C)]^\lambda [\text{vol}(D)]^{1-\lambda} \qquad \text{for all } \lambda \in [0,1]. \tag{3.45}$$

This form of the Brunn–Minkowski inequality can be used to establish Lévy-type concentration for the uniform measure on the sphere, albeit with slightly weaker constants than the derivation in Example 3.10. In Exercise 3.10, we explore the equivalence between inequality (3.45) and our original statement (3.43) of the Brunn–Minkowski inequality.

The modified form (3.45) of the Brunn–Minkowski inequality also leads naturally to a functional-analytic generalization, due to Prékopa and Leindler. In turn, this generalized inequality can be used to derive concentration inequalities for strongly log-concave measures.

Theorem 3.15 (Prékopa–Leindler inequality) *Let u, v, w be non-negative integrable functions such that, for some $\lambda \in [0,1]$, we have*

$$w(\lambda x + (1-\lambda)y) \geq [u(x)]^\lambda [v(y)]^{1-\lambda} \qquad \textit{for all } x, y \in \mathbb{R}^n. \tag{3.46}$$

Then

$$\int w(x)\, dx \geq \left(\int u(x)\, dx\right)^\lambda \left(\int v(x)\, dx\right)^{1-\lambda}. \tag{3.47}$$

https://doi.org/10.1017/9781108627771.003 Published online by Cambridge University Press

In order to see how this claim implies the classical Brunn–Minkowski inequality (3.45), consider the choices

$$u(x) = \mathbb{I}_C(x), \quad v(x) = \mathbb{I}_D(x) \quad \text{and} \quad w(x) = \mathbb{I}_{\lambda C+(1-\lambda)D}(x),$$

respectively. Here $\mathbb{I}_C$ denotes the binary-valued indicator function for the event $\{x \in C\}$, with the other indicators defined in an analogous way. In order to show that the classical inequality (3.45) follows as a consequence of Theorem 3.15, we need to verify that

$$\mathbb{I}_{\lambda C+(1-\lambda)D}(\lambda x + (1-\lambda)y) \geq [\mathbb{I}_C(x)]^{\lambda}[\mathbb{I}_D(y)]^{1-\lambda} \qquad \text{for all } x, y \in \mathbb{R}^n.$$

For $\lambda = 0$ or $\lambda = 1$, the claim is immediate. For any $\lambda \in (0, 1)$, if either $x \notin C$ or $y \notin D$, the right-hand side is zero, so the statement is trivial. Otherwise, if $x \in C$ and $y \in D$, then both sides are equal to one.

The Prékopa–Leindler inequality can be used to establish some interesting concentration inequalities of Lipschitz functions for a particular subclass of distributions, one which allows for some dependence. In particular, we say that a distribution $\mathbb{P}$ with a density p (with respect to the Lebesgue measure) is a *strongly log-concave distribution* if the function $\log p$ is strongly concave. Equivalently stated, this condition means that the density can be written in the form $p(x) = \exp(-\psi(x))$, where the function $\psi \colon \mathbb{R}^n \to \mathbb{R}$ is strongly convex, meaning that there is some $\gamma > 0$ such that

$$\lambda\psi(x) + (1-\lambda)\psi(y) - \psi(\lambda x + (1-\lambda)y) \geq \frac{\gamma}{2}\lambda(1-\lambda)\,\|x-y\|_2^2 \tag{3.48}$$

for all $\lambda \in [0, 1]$, and $x, y \in \mathbb{R}^n$. For instance, it is easy to verify that the distribution of a standard Gaussian vector in n dimensions is strongly log-concave with parameter $\gamma = 1$. More generally, any Gaussian distribution with covariance matrix $\mathbf{\Sigma} > 0$ is strongly log-concave with parameter $\gamma = \gamma_{\min}(\mathbf{\Sigma}^{-1}) = (\gamma_{\max}(\mathbf{\Sigma}))^{-1}$. In addition, there are a variety of non-Gaussian distributions that are also strongly log-concave. For any such distribution, Lipschitz functions are guaranteed to concentrate, as summarized in the following:

Theorem 3.16 *Let $\mathbb{P}$ be any strongly log-concave distribution with parameter $\gamma > 0$. Then for any function $f \colon \mathbb{R}^n \to \mathbb{R}$ that is L-Lipschitz with respect to Euclidean norm, we have*

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2e^{-\frac{\gamma t^2}{4L^2}}. \tag{3.49}$$

Remark: Since the standard Gaussian distribution is log-concave with parameter $\gamma = 1$, this theorem implies our earlier result (Theorem 2.26), albeit with a sub-optimal constant in the exponent.

Proof Let h be an arbitrary zero-mean function with Lipschitz constant L with respect to the Euclidean norm. It suffices to show that $\mathbb{E}[e^{h(X)}] \leq e^{\frac{L^2}{\gamma}}$. Indeed, if this inequality holds, then, given an arbitrary function f with Lipschitz constant K and $\lambda \in \mathbb{R}$, we can apply

https://doi.org/10.1017/9781108627771.003 Published online by Cambridge University Press

this inequality to the zero-mean function $h := \lambda(f - \mathbb{E}[f(X)])$, which has Lipschitz constant $L = \lambda K$. Doing so yields the bound

$$\mathbb{E}[e^{\lambda(f(X)-\mathbb{E}[f(X)])}] \leq e^{\frac{\lambda^2 K^2}{\gamma}} \qquad \text{for all } \lambda \in \mathbb{R},$$

which shows that $f(X) - \mathbb{E}[f(X)]$ is a sub-Gaussian random variable. As shown in Chapter 2, this type of uniform control on the moment generating function implies the claimed tail bound.

Accordingly, for a given zero-mean function h that is L-Lipschitz and for given $\lambda \in (0, 1)$ and $x, y \in \mathbb{R}^n$, define the function

$$g(y) := \inf_{x \in \mathbb{R}^n} \left\{ h(x) + \frac{\gamma}{4} \|x - y\|_2^2 \right\},$$

known as the inf-convolution of h with the rescaled Euclidean norm. With this definition, the proof is based on applying the Prékopa–Leindler inequality with $\lambda = 1/2$ to the triplet of functions $w(z) \equiv p(z) = \exp(-\psi(z))$, the density of $\mathbb{P}$, and the pair of functions

$$u(x) := \exp(-h(x) - \psi(x)) \quad \text{and} \quad v(y) := \exp(g(y) - \psi(y)).$$

We first need to verify that the inequality (3.46) holds with $\lambda = 1/2$. By the definitions of u and v, the logarithm of the right-hand side of inequality (3.46)—call it R for short—is given by

$$R = \tfrac{1}{2}\{g(y) - h(x)\} - \tfrac{1}{2}\psi(x) - \tfrac{1}{2}\psi(y) = \tfrac{1}{2}\{g(y) - h(x) - 2E(x, y)\} - \psi(x/2 + y/2),$$

where $E(x, y) := \frac{1}{2}\psi(x) + \frac{1}{2}\psi(y) - \psi(x/2 + y/2)$. Since $\mathbb{P}$ is a γ-log-concave distribution, the function ψ is γ strongly convex, and hence $2E(x, y) \geq \frac{\gamma}{4} \|x - y\|_2^2$. Substituting into the earlier representation of R, we find that

$$R \leq \frac{1}{2} \left\{ g(y) - h(x) - \frac{\gamma}{4} \|x - y\|_2^2 \right\} - \psi(x/2 + y/2) \leq -\psi(x/2 + y/2),$$

where the final inequality follows from the definition of the inf-convolution g. We have thus verified condition (3.46) with $\lambda = 1/2$.

Now since $\int w(x)\, dx = \int p(x)\, dx = 1$ by construction, the Prékopa–Leindler inequality implies that

$$0 \geq \frac{1}{2} \log \int e^{-h(x)-\psi(x)}\, dx + \frac{1}{2} \log \int e^{g(y)-\psi(y)}\, dy.$$

Rewriting the integrals as expectations and rearranging yields

$$\mathbb{E}[e^{g(Y)}] \leq \frac{1}{\mathbb{E}[e^{-h(X)}]} \overset{\text{(i)}}{\leq} \frac{1}{e^{\mathbb{E}[-h(X)]}} \overset{\text{(ii)}}{=} 1, \tag{3.50}$$

where step (i) follows from Jensen's inequality, and convexity of the function $t \mapsto \exp(-t)$, and step (ii) uses the fact that $\mathbb{E}[-h(X)] = 0$ by assumption. Finally, since h is an L-Lipschitz

https://doi.org/10.1017/9781108627771.003 Published online by Cambridge University Press

function, we have $|h(x) - h(y)| \leq L\,\|x - y\|_2$, and hence

$$g(y) = \inf_{x\in\mathbb{R}^n}\left\{h(x) + \frac{\gamma}{4}\,\|x-y\|_2^2\right\} \geq h(y) + \inf_{x\in\mathbb{R}^n}\left\{-L\,\|x-y\|_2 + \frac{\gamma}{4}\,\|x-y\|_2^2\right\}$$
$$= h(y) - \frac{L^2}{\gamma}.$$

Combined with the bound (3.50), we conclude that $\mathbb{E}[e^{h(Y)}] \leq \exp(\frac{L^2}{\gamma})$, as claimed. □

3.3 Wasserstein distances and information inequalities

We now turn to the topic of Wasserstein distances and information inequalities, also known as *transportation cost inequalities*. On one hand, the transportation cost approach can be used to obtain some sharp results for Lipschitz functions of independent random variables. Perhaps more importantly, it is especially well suited to certain types of dependent random variables, such as those arising in Markov chains and other types of mixing processes.

3.3.1 Wasserstein distances

We begin by defining the notion of a Wasserstein distance. Given a metric space $(\mathcal{X},\rho)$, a function $f\colon \mathcal{X}\to\mathbb{R}$ is L-Lipschitz with respect to the metric ρ if

$$|f(x) - f(x')| \leq L\rho(x,x') \qquad \text{for all } x, x' \in \mathcal{X}, \tag{3.51}$$

and we use $\|f\|_{\mathrm{Lip}}$ to denote the smallest L for which this inequality holds. Given two probability distributions $\mathbb{Q}$ and $\mathbb{P}$ on $\mathcal{X}$, we can then measure the distance between them via

$$W_\rho(\mathbb{Q},\,\mathbb{P}) = \sup_{\|f\|_{\mathrm{Lip}}\leq 1}\left[\int f\,d\mathbb{Q} - \int f\,d\mathbb{P}\right], \tag{3.52}$$

where the supremum ranges over all 1-Lipschitz functions. This distance measure is referred to as the *Wasserstein metric induced by* ρ. It can be verified that, for each choice of the metric ρ, this definition defines a distance on the space of probability measures.

Example 3.17 (Hamming metric and total variation distance) Consider the Hamming metric $\rho(x,x') = \mathbb{I}[x \neq x']$. We claim that, in this case, the associated Wasserstein distance is equivalent to the *total variation distance*

$$\|\mathbb{Q} - \mathbb{P}\|_{\mathrm{TV}} := \sup_{A\subseteq\mathcal{X}}|\mathbb{Q}(A) - \mathbb{P}(A)|, \tag{3.53}$$

where the supremum ranges over all measurable subsets A. To see this equivalence, note that any function that is 1-Lipschitz with respect to the Hamming distance satisfies the bound $|f(x) - f(x')| \leq 1$. Since the supremum (3.52) is invariant to constant offsets of the function, we may restrict the supremum to functions such that $f(x) \in [0,1]$ for all $x \in \mathcal{X}$, thereby obtaining

$$W_{\mathrm{Ham}}(\mathbb{Q},\,\mathbb{P}) = \sup_{f\colon \mathcal{X}\to[0,1]}\int f\,(d\mathbb{Q} - d\mathbb{P}) \overset{\text{(i)}}{=} \|\mathbb{Q} - \mathbb{P}\|_{\mathrm{TV}},$$

https://doi.org/10.1017/9781108627771.003 Published online by Cambridge University Press

where equality (i) follows from Exercise 3.13.

In terms of the underlying densities[3] p and q taken with respect to a base measure ν, we can write

$$W_{\text{Ham}}(\mathbb{Q}, \mathbb{P}) = \|\mathbb{Q} - \mathbb{P}\|_{\text{TV}} = \frac{1}{2}\int |p(x) - q(x)|\nu(dx),$$

corresponding to (one half) the $L^1(\nu)$-norm between the densities. Again, see Exercise 3.13 for further details on this equivalence. ♣

By a classical and deep result in duality theory (see the bibliographic section for details), any Wasserstein distance has an equivalent definition as a type of coupling-based distance. A distribution $\mathbb{M}$ on the product space $\mathcal{X} \otimes \mathcal{X}$ is a *coupling* of the pair $(\mathbb{Q}, \mathbb{P})$ if its marginal distributions in the first and second coordinates coincide with $\mathbb{Q}$ and $\mathbb{P}$, respectively. In order to see the relation to the Wasserstein distance, let $f\colon \mathcal{X} \to \mathbb{R}$ be any 1-Lipschitz function, and let $\mathbb{M}$ be any coupling. We then have

$$\int \rho(x, x')\, d\mathbb{M}(x, x') \overset{\text{(i)}}{\geq} \int (f(x) - f(x'))\, d\mathbb{M}(x, x') \overset{\text{(ii)}}{=} \int f\,(d\mathbb{P} - d\mathbb{Q}), \tag{3.54}$$

where the inequality (i) follows from the 1-Lipschitz nature of f, and the equality (ii) follows since $\mathbb{M}$ is a coupling. The *Kantorovich–Rubinstein duality* guarantees the following important fact: if we minimize over all possible couplings, then this argument can be reversed, and in fact we have the equivalence

$$\underbrace{\sup_{\|f\|_{\text{Lip}}\leq 1} \int f\,(d\mathbb{Q} - d\mathbb{P})}_{W_\rho(\mathbb{P},\mathbb{Q})} = \inf_{\mathbb{M}} \int_{\mathcal{X}\times\mathcal{X}} \rho(x, x')\, d\mathbb{M}(x, x') = \inf_{\mathbb{M}} \mathbb{E}_{\mathbb{M}}[\rho(X, X')], \tag{3.55}$$

where the infimum ranges over all couplings $\mathbb{M}$ of the pair $(\mathbb{P}, \mathbb{Q})$. This coupling-based representation of the Wasserstein distance plays an important role in many of the proofs to follow.

The term "transportation cost" arises from the following interpretation of coupling-based representation (3.55). For concreteness, let us consider the case where $\mathbb{P}$ and $\mathbb{Q}$ have densities p and q with respect to Lebesgue measure on $\mathcal{X}$, and the coupling $\mathbb{M}$ has density m with respect to Lebesgue measure on the product space. The density p can be viewed as describing some initial distribution of mass over the space $\mathcal{X}$, whereas the density q can be interpreted as some desired distribution of the mass. Our goal is to shift mass so as to transform the initial distribution p to the desired distribution q. The quantity $\rho(x, x')\, dx\, dx'$ can be interpreted as the cost of transporting a small increment of mass dx to the new increment dx'. The joint distribution $m(x, x')$ is known as a *transportation plan*, meaning a scheme for shifting mass so that p is transformed to q. Combining these ingredients, we conclude that the transportation cost associated with the plan m is given by

$$\int_{\mathcal{X}\times\mathcal{X}} \rho(x, x') m(x, x')\, dx\, dx',$$

and minimizing over all admissible plans—that is, those that marginalize down to p and q,

[3] This assumption entails no loss of generality, since $\mathbb{P}$ and $\mathbb{Q}$ both have densities with respect to $\nu = \frac{1}{2}(\mathbb{P} + \mathbb{Q})$.

https://doi.org/10.1017/9781108627771.003 Published online by Cambridge University Press

respectively—yields the Wasserstein distance.

3.3.2 Transportation cost and concentration inequalities

Let us now turn to the notion of a transportation cost inequality, and its implications for the concentration of measure. Transportation cost inequalities are based on upper bounding the Wasserstein distance $W_\rho(\mathbb{Q}, \mathbb{P})$ in terms of the *Kullback–Leibler (KL) divergence*. Given two distributions $\mathbb{Q}$ and $\mathbb{P}$, the KL divergence between them is given by

$$D(\mathbb{Q} \,\|\, \mathbb{P}) := \begin{cases} \mathbb{E}_{\mathbb{Q}}\Big[\log \frac{d\mathbb{Q}}{d\mathbb{P}}\Big] & \text{when } \mathbb{Q} \text{ is absolutely continuous with respect to } \mathbb{P}, \\ +\infty & \text{otherwise.} \end{cases} \tag{3.56}$$

If the measures have densities[4] with respect to some underlying measure ν—say q and p—then the Kullback–Leibler divergence can be written in the form

$$D(\mathbb{Q} \,\|\, \mathbb{P}) = \int_{\mathcal{X}} q(x) \log \frac{q(x)}{p(x)} \nu(dx). \tag{3.57}$$

Although the KL divergence provides a measure of distance between distributions, it is not actually a metric (since, for instance, it is not symmetric in general).

We say that a transportation cost inequality is satisfied when the Wasserstein distance is upper bounded by a multiple of the square-root KL divergence.

Definition 3.18 For a given metric ρ, the probability measure $\mathbb{P}$ is said to satisfy a *ρ-transportation cost inequality* with parameter $\gamma > 0$ if

$$W_\rho(\mathbb{Q}, \mathbb{P}) \leq \sqrt{2\gamma D(\mathbb{Q} \,\|\, \mathbb{P})} \tag{3.58}$$

for all probability measures $\mathbb{Q}$.

Such results are also known as *information inequalities*, due to the role of the Kullback–Leibler divergence in information theory. A classical example of an information inequality is the *Pinsker–Csiszár–Kullback inequality*, which relates the total variation distance with the KL divergence. More precisely, for all probability distributions $\mathbb{P}$ and $\mathbb{Q}$, we have

$$\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} \leq \sqrt{\tfrac{1}{2} D(\mathbb{Q} \,\|\, \mathbb{P})}. \tag{3.59}$$

From our development in Example 3.17, this inequality corresponds to a transportation cost inequality, in which $\gamma = 1/4$ and the Wasserstein distance is based on the Hamming norm $\rho(x, x') = \mathbb{I}[x \neq x']$. As will be seen shortly, this inequality can be used to recover the bounded differences inequality, corresponding to a concentration statement for functions that are Lipschitz with respect to the Hamming norm. See Exercise 15.6 in Chapter 15 for

[4] In the special case of a discrete space $\mathcal{X}$, and probability mass functions q and p, we have $D(\mathbb{Q} \,\|\, \mathbb{P}) = \sum_{x \in \mathcal{X}} q(x) \log \frac{q(x)}{p(x)}$.

https://doi.org/10.1017/9781108627771.003 Published online by Cambridge University Press

the proof of this bound.

By the definition (3.52) of the Wasserstein distance, the transportation cost inequality (3.58) can be used to upper bound the deviation $\int f\, d\mathbb{Q} - \int f\, d\mathbb{P}$ in terms of the Kullback–Leibler divergence $D(\mathbb{Q} \,\|\, \mathbb{P})$. As shown by the following result, a particular choice of distribution $\mathbb{Q}$ can be used to derive a concentration bound for f under $\mathbb{P}$. In this way, a transportation cost inequality leads to concentration bounds for Lipschitz functions:

Theorem 3.19 (From transportation cost to concentration) *Consider a metric measure space* $(\mathbb{P}, \mathcal{X}, \rho)$*, and suppose that* $\mathbb{P}$ *satisfies the* ρ*-transportation cost inequality* (3.58). *Then its concentration function satisfies the bound*

$$\alpha_{\mathbb{P},(\mathcal{X},\rho)}(t) \leq 2 \exp\left(-\frac{t^2}{2\gamma}\right). \tag{3.60}$$

Moreover, for any $X \sim \mathbb{P}$ *and any L-Lipschitz function* $f \colon \mathcal{X} \to \mathbb{R}$*, we have the concentration inequality*

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2 \exp\left(-\frac{t^2}{2\gamma L^2}\right). \tag{3.61}$$

Remarks: By Proposition 3.11, the bound (3.60) implies that

$$\mathbb{P}[|f(X) - m_f| \geq t] \leq 2 \exp\left(-\frac{t^2}{2\gamma L^2}\right), \tag{3.62}$$

where m_f is any median of f. In turn, this bound can be used to establish concentration around the mean, albeit with worse constants than the bound (3.61). (See Exercise 2.14 for details on this equivalence.) In our proof, we make use of separate arguments for the median and mean, so as to obtain sharp constants.

Proof We begin by proving the bound (3.60). For any set A with $\mathbb{P}[A] \geq 1/2$ and a given $\epsilon > 0$, consider the set

$$B := (A^\epsilon)^c = \{y \in \mathcal{X} \mid \rho(x, y) \geq \epsilon \quad \forall\, x \in A\}.$$

If $\mathbb{P}(A^\epsilon) = 1$, then the proof is complete, so that we may assume that $\mathbb{P}(B) > 0$.

By construction, we have $\rho(A, B) := \inf_{x \in A} \inf_{y \in B} \rho(x, y) \geq \epsilon$. On the other hand, let $\mathbb{P}_A$ and $\mathbb{P}_B$ denote the distributions of $\mathbb{P}$ conditioned on A and B, and let $\mathbb{M}$ denote any coupling of this pair. Since the marginals of $\mathbb{M}$ are supported on A and B, respectively, we have $\rho(A, B) \leq \int \rho(x, x')\, d\mathbb{M}(x, x')$. Taking the infimum over all couplings, we conclude that $\epsilon \leq \rho(A, B) \leq W_\rho(\mathbb{P}_A,\, \mathbb{P}_B)$.

Now applying the triangle inequality, we have

$$\begin{aligned} \epsilon \leq W_\rho(\mathbb{P}_A,\, \mathbb{P}_B) \leq W_\rho(\mathbb{P},\, \mathbb{P}_A) + W_\rho(\mathbb{P},\, \mathbb{P}_B) &\overset{\text{(ii)}}{\leq} \sqrt{\gamma D(\mathbb{P}_A \,\|\, \mathbb{P})} + \sqrt{\gamma D(\mathbb{P}_B \,\|\, \mathbb{P})} \\ &\overset{\text{(iii)}}{\leq} \sqrt{2\gamma}\, \{D(\mathbb{P}_A \,\|\, \mathbb{P}) + D(\mathbb{P}_B \,\|\, \mathbb{P})\}^{1/2}, \end{aligned}$$

https://doi.org/10.1017/9781108627771.003 Published online by Cambridge University Press

where step (ii) follows from the transportation cost inequality, and step (iii) follows from the inequality $(a+b)^2 \leq 2a^2 + 2b^2$.

It remains to compute the Kullback–Leibler divergences. For any measurable set C, we have $\mathbb{P}_A(C) = \mathbb{P}(C \cap A)/\mathbb{P}(A)$, so that $D(\mathbb{P}_A \,\|\, \mathbb{P}) = \log \frac{1}{\mathbb{P}(A)}$. Similarly, we have $D(\mathbb{P}_B \,\|\, \mathbb{P}) = \log \frac{1}{\mathbb{P}(B)}$. Combining the pieces, we conclude that

$$\epsilon^2 \leq 2\gamma\{\log(1/\mathbb{P}(A)) + \log(1/\mathbb{P}(B))\} = 2\gamma \log\left(\frac{1}{\mathbb{P}(A)\mathbb{P}(B)}\right),$$

or equivalently $\mathbb{P}(A)\mathbb{P}(B) \leq \exp\left(-\frac{\epsilon^2}{2\gamma}\right)$. Since $\mathbb{P}(A) \geq 1/2$ and $B = (A^\epsilon)^c$, we conclude that $\mathbb{P}(A^\epsilon) \geq 1 - 2\exp\left(-\frac{\epsilon^2}{2\gamma}\right)$. Since A was an arbitrary set with $\mathbb{P}(A) \geq 1/2$, the bound (3.60) follows.

We now turn to the proof of the concentration statement (3.61) for the mean. If one is not concerned about constants, such a bound follows immediately by combining claim (3.60) with the result of Exercise 2.14. Here we present an alternative proof with the dual goals of obtaining the sharp result and illustrating a different proof technique. Throughout this proof, we use $\mathbb{E}_{\mathbb{Q}}[f]$ and $\mathbb{E}_{\mathbb{P}}[f]$ to denote the mean of the random variable $f(X)$ when $X \sim \mathbb{Q}$ and $X \sim \mathbb{P}$, respectively. We begin by observing that

$$\int f\,(d\mathbb{Q} - d\mathbb{P}) \overset{\text{(i)}}{\leq} L W_\rho(\mathbb{Q},\, \mathbb{P}) \overset{\text{(ii)}}{\leq} \sqrt{2L^2\gamma D(\mathbb{Q} \,\|\, \mathbb{P})},$$

where step (i) follows from the L-Lipschitz condition on f and the definition (3.52); and step (ii) follows from the information inequality (3.58). For any positive numbers (u, v, λ), we have $\sqrt{2uv} \leq \frac{u}{2}\lambda + \frac{v}{\lambda}$. Applying this inequality with $u = L^2\gamma$ and $v = D(\mathbb{Q} \,\|\, \mathbb{P})$ yields

$$\int f\,(d\mathbb{Q} - d\mathbb{P}) \leq \frac{\lambda\gamma L^2}{2} + \frac{1}{\lambda} D(\mathbb{Q} \,\|\, \mathbb{P}), \tag{3.63}$$

valid for all $\lambda > 0$.

Now define a distribution $\mathbb{Q}$ with Radon–Nikodym derivative $\frac{d\mathbb{Q}}{d\mathbb{P}}(x) = e^{g(x)}/\mathbb{E}_{\mathbb{P}}[e^{g(X)}]$, where $g(x) := \lambda(f(x) - \mathbb{E}_{\mathbb{P}}(f)) - \frac{L^2\gamma\lambda^2}{2}$. (Note that our proof of the bound (3.61) ensures that $\mathbb{E}_{\mathbb{P}}[e^{g(X)}]$ exists.) With this choice, we have

$$D(\mathbb{Q} \,\|\, \mathbb{P}) = \mathbb{E}_{\mathbb{Q}} \log\left(\frac{e^{g(X)}}{\mathbb{E}_{\mathbb{P}}[e^{g(X)}]}\right) = \lambda\{\mathbb{E}_{\mathbb{Q}}(f(X)) - \mathbb{E}_{\mathbb{P}}(f(X))\} - \frac{\gamma L^2 \lambda^2}{2} - \log \mathbb{E}_{\mathbb{P}}[e^{g(X)}].$$

Combining with inequality (3.63) and performing some algebra (during which the reader should recall that $\lambda > 0$), we find that $\log \mathbb{E}_{\mathbb{P}}[e^{g(X)}] \leq 0$, or equivalently

$$\mathbb{E}_{\mathbb{P}}[e^{\lambda(f(X) - \mathbb{E}_{\mathbb{P}}[f(X')])}] \leq e^{\frac{\lambda^2 \gamma L^2}{2}}.$$

The upper tail bound thus follows by the Chernoff bound. The same argument can be applied to $-f$, which yields the lower tail bound. □

3.3.3 Tensorization for transportation cost

Based on Theorem 3.19, we see that transportation cost inequalities can be translated into concentration inequalities. Like entropy, transportation cost inequalities behave nicely for product measures, and can be combined in an additive manner. Doing so yields concentration inequalities for Lipschitz functions in the higher-dimensional space. We summarize in the following:

Proposition 3.20 *Suppose that, for each $k = 1, 2, \ldots, n$, the univariate distribution $\mathbb{P}_k$ satisfies a ρ_k-transportation cost inequality with parameter γ_k. Then the product distribution $\mathbb{P} = \bigotimes_{k=1}^n \mathbb{P}_k$ satisfies the transportation cost inequality*

$$W_\rho(\mathbb{Q}, \mathbb{P}) \leq \sqrt{2\Big(\sum_{k=1}^n \gamma_k\Big) D(\mathbb{Q} \,\|\, \mathbb{P})} \qquad \textit{for all distributions } \mathbb{Q}, \tag{3.64}$$

where the Wasserstein metric is defined using the distance $\rho(x, y) := \sum_{k=1}^n \rho_k(x_k, y_k)$.

Before turning to the proof of Proposition 3.20, it is instructive to see how, in conjunction with Theorem 3.19, it can be used to recover the bounded differences inequality.

Example 3.21 (Bounded differences inequality) Suppose that f satisfies the bounded differences inequality with parameter L_k in coordinate k. Then using the triangle inequality and the bounded differences property, it can be verified that f is a 1-Lipschitz function with respect to the rescaled Hamming metric

$$\rho(x, y) := \sum_{k=1}^n \rho_k(x_k, y_k), \qquad \text{where } \rho_k(x_k, y_k) := L_k \, \mathbb{I}[x_k \neq y_k].$$

By the Pinsker–Csiszár–Kullback inequality (3.59), each univariate distribution $\mathbb{P}_k$ satisfies a ρ_k-transportation cost inequality with parameter $\gamma_k = \frac{L_k^2}{4}$, so that Proposition 3.20 implies that $\mathbb{P} = \bigotimes_{k=1}^n \mathbb{P}_k$ satisfies a ρ-transportation cost inequality with parameter $\gamma := \frac{1}{4} \sum_{k=1}^n L_k^2$. Since f is 1-Lipschitz with respect to the metric ρ, Theorem 3.19 implies that

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2 \exp\left(-\frac{2t^2}{\sum_{k=1}^n L_k^2}\right). \tag{3.65}$$

In this way, we recover the bounded differences inequality from Chapter 2 from a transportation cost argument. ♣

Our proof of Proposition 3.20 is based on the coupling-based characterization (3.55) of Wasserstein distances.

Proof Letting $\mathbb{Q}$ be an arbitrary distribution over the product space $\mathcal{X}^n$, we construct a coupling $\mathbb{M}$ of the pair $(\mathbb{P}, \mathbb{Q})$. For each $j = 2, \ldots, n$, let $\mathbb{M}_1^j$ denote the joint distribution over the pair $(X_1^j, Y_1^j) = (X_1, \ldots, X_j, Y_1, \ldots, Y_j)$, and let $\mathbb{M}_{j|j-1}$ denote the conditional distribution of (X_j, Y_j) given (X_1^{j-1}, Y_1^{j-1}). By the dual representation (3.55), we have

$$W_\rho(\mathbb{Q}, \mathbb{P}) \leq \mathbb{E}_{\mathbb{M}_1}[\rho_1(X_1, Y_1)] + \sum_{j=2}^n \mathbb{E}_{\mathbb{M}_1^{j-1}}[\mathbb{E}_{\mathbb{M}_{j|j-1}}[\rho_j(X_j, Y_j)]],$$

https://doi.org/10.1017/9781108627771.003 Published online by Cambridge University Press

where $\mathbb{M}_j$ denotes the marginal distribution over the pair (X_j, Y_j). We now define our coupling $\mathbb{M}$ in an inductive manner as follows. First, choose $\mathbb{M}_1$ to be an optimal coupling of the pair $(\mathbb{P}_1, \mathbb{Q}_1)$, thereby ensuring that

$$\mathbb{E}_{\mathbb{M}_1}[\rho_1(X_1, Y_1)] \overset{\text{(i)}}{=} W_\rho(\mathbb{Q}_1, \mathbb{P}_1) \overset{\text{(ii)}}{\leq} \sqrt{2\gamma_1 D(\mathbb{Q}_1 \,\|\, \mathbb{P}_1)},$$

where equality (i) follows by the optimality of the coupling, and inequality (ii) follows from the assumed transportation cost inequality for $\mathbb{P}_1$. Now assume that the joint distribution over (X_1^{j-1}, Y_1^{j-1}) has been defined. We choose conditional distribution $\mathbb{M}_{j|j-1}(\cdot \mid x_1^{j-1}, y_1^{j-1})$ to be an optimal coupling for the pair $(\mathbb{P}_j, \mathbb{Q}_{j|j-1}(\cdot \mid y_1^{j-1}))$, thereby ensuring that

$$\mathbb{E}_{\mathbb{M}_{j|j-1}}[\rho_j(X_j, Y_j)] \leq \sqrt{2\gamma_j D(\mathbb{Q}_{j|j-1}(\cdot \mid y_1^{j-1}) \,\|\, \mathbb{P}_j)},$$

valid for each y_1^{j-1}. Taking averages over Y_1^{j-1} with respect to the marginal distribution $\mathbb{M}_1^{j-1}$—or, equivalently, the marginal $\mathbb{Q}_1^{j-1}$—the concavity of the square-root function and Jensen's inequality implies that

$$\mathbb{E}_{\mathbb{M}_1^{j-1}}[\mathbb{E}_{\mathbb{M}_{j|j-1}}[\rho_j(X_j, Y_j)]] \leq \sqrt{2\gamma_j \mathbb{E}_{\mathbb{Q}_1^{j-1}} D(\mathbb{Q}_{j|j-1}(\cdot \mid Y_1^{j-1}) \,\|\, \mathbb{P}_j)}.$$

Combining the ingredients, we obtain

$$\begin{aligned} W_\rho(\mathbb{Q}, \mathbb{P}) &\leq \sqrt{2\gamma_1 D(\mathbb{Q}_1 \,\|\, \mathbb{P}_1)} + \sum_{j=2}^{n} \sqrt{2\gamma_j \mathbb{E}_{\mathbb{Q}_1^{j-1}}[D(\mathbb{Q}_{j|j-1}(\cdot \mid Y_1^{j-1}) \,\|\, \mathbb{P}_j)]} \\ &\overset{\text{(i)}}{\leq} \sqrt{2\Big(\sum_{j=1}^{n} \gamma_j\Big)} \sqrt{D(\mathbb{Q}_1 \,\|\, \mathbb{P}_1) + \sum_{j=2}^{n} \mathbb{E}_{\mathbb{Q}_1^{j-1}}[D(\mathbb{Q}_{j|j-1}(\cdot \mid Y_1^{j-1}) \,\|\, \mathbb{P}_j)]} \\ &\overset{\text{(ii)}}{=} \sqrt{2\Big(\sum_{j=1}^{n} \gamma_j\Big) D(\mathbb{Q} \,\|\, \mathbb{P})}, \end{aligned}$$

where step (i) by follows the Cauchy–Schwarz inequality, and equality (ii) uses the chain rule for Kullback–Leibler divergence from Exercise 3.2. □

In Exercise 3.14, we sketch out an alternative proof of Proposition 3.20, one which makes direct use of the Lipschitz characterization of the Wasserstein distance.

3.3.4 Transportation cost inequalities for Markov chains

As mentioned previously, the transportation cost approach has some desirable features in application to Lipschitz functions involving certain types of dependent random variables. Here we illustrate this type of argument for the case of a Markov chain. (See the bibliographic section for references to more general results on concentration for dependent random variables.)

More concretely, let $(X_1, \ldots, X_n)$ be a random vector generated by a Markov chain, where each X_i takes values in a countable space $\mathcal{X}$. Its distribution $\mathbb{P}$ over $\mathcal{X}^n$ is defined by an initial distribution $X_1 \sim \mathbb{P}_1$, and the transition kernels

$$\mathbb{K}_{i+1}(x_{i+1} \mid x_i) = \mathbb{P}_{i+1}(X_{i+1} = x_{i+1} \mid X_i = x_i). \tag{3.66}$$

Here we focus on discrete state Markov chains that are β-contractive, meaning that there exists some $\beta \in [0, 1)$ such that

$$\max_{i=1,\ldots,n-1} \sup_{x_i, x_i'} \|\mathbb{K}_{i+1}(\cdot \mid x_i) - \mathbb{K}_{i+1}(\cdot \mid x_i')\|_{\mathrm{TV}} \leq \beta, \tag{3.67}$$

where the total variation norm (3.53) was previously defined.

Theorem 3.22 *Let $\mathbb{P}$ be the distribution of a β-contractive Markov chain* (3.67) *over the discrete space $\mathcal{X}^n$. Then for any other distribution $\mathbb{Q}$ over $\mathcal{X}^n$, we have*

$$W_\rho(\mathbb{Q}, \mathbb{P}) \leq \frac{1}{1-\beta} \sqrt{\frac{n}{2} D(\mathbb{Q} \,\|\, \mathbb{P})}, \tag{3.68}$$

where the Wasserstein distance is defined with respect to the Hamming norm $\rho(x, y) = \sum_{i=1}^n \mathbb{I}[x_i \neq y_i]$.

Remark: See the bibliography section for references to proofs of this result. Using Theorem 3.19, an immediate corollary of the bound (3.68) is that for any function $f \colon \mathcal{X}^n \to \mathbb{R}$ that is L-Lipschitz with respect to the Hamming norm, we have

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2 \exp\left(-\frac{2(1-\beta)^2 t^2}{nL^2}\right). \tag{3.69}$$

Note that this result is a strict generalization of the bounded difference inequality for independent random variables, to which it reduces when $\beta = 0$.

Example 3.23 (Parameter estimation for a binary Markov chain) Consider a Markov chain over binary variables $X_i \in \{0, 1\}^2$ specified by an initial distribution $\mathbb{P}_1$ that is uniform, and the transition kernel

$$\mathbb{K}_{i+1}(x_{i+1} \mid x_i) = \begin{cases} \frac{1}{2}(1+\delta) & \text{if } x_{i+1} = x_i, \\ \frac{1}{2}(1-\delta) & \text{if } x_{i+1} \neq x_i, \end{cases}$$

where $\delta \in [0, 1]$ is a "stickiness" parameter. Suppose that our goal is to estimate the parameter δ based on an n-length vector $(X_1, \ldots, X_n)$ drawn according to this chain. An unbiased estimate of $\frac{1}{2}(1+\delta)$ is given by the function

$$f(X_1, \ldots, X_n) := \frac{1}{n-1} \sum_{i=1}^{n-1} \mathbb{I}[X_i = X_{i+1}],$$

corresponding to the fraction of times that successive samples take the same value. We claim that f satisfies the concentration inequality

$$\mathbb{P}[|f(X) - \tfrac{1}{2}(1+\delta)| \geq t] \leq 2e^{-\frac{(n-1)^2(1-\delta)^2 t^2}{2n}} \leq 2e^{-\frac{(n-1)(1-\delta)^2 t^2}{4}}. \tag{3.70}$$

Following some calculation, we find that the chain is β-contractive with $\beta = \delta$. Moreover, the function f is $\frac{2}{n-1}$-Lipschitz with respect to the Hamming norm. Consequently, the bound (3.70) follows as a consequence of our earlier general result (3.69). ♣

3.3.5 Asymmetric coupling cost

Thus far, we have considered various types of Wasserstein distances, which can be used to obtain concentration for Lipschitz functions. However, this approach—as with most methods that involve Lipschitz conditions with respect to ℓ_1-type norms—typically does not yield dimension-independent bounds. By contrast, as we have seen previously, Lipschitz conditions based on the ℓ_2-norm often do lead to dimension-independent results.

With this motivation in mind, this section is devoted to consideration of another type of coupling-based distance between probability distributions, but one that is asymmetric in its two arguments, and of a quadratic nature. In particular, we define

$$C(\mathbb{Q}, \mathbb{P}) := \inf_{\mathbb{M}} \sqrt{\int \sum_{i=1}^{n} (\mathbb{M}[Y_i \neq x_i \mid X_i = x_i])^2 \, d\mathbb{P}(x)}, \tag{3.71}$$

where once again the infimum ranges over all couplings $\mathbb{M}$ of the pair $(\mathbb{P}, \mathbb{Q})$. This distance is relatively closely related to the total variation distance; in particular, it can be shown that an equivalent representation for this asymmetric distance is

$$C(\mathbb{Q}, \mathbb{P}) = \sqrt{\int \left|1 - \frac{d\mathbb{Q}}{d\mathbb{P}}(x)\right|_+^2 d\mathbb{P}(x)}, \tag{3.72}$$

where $t_+ := \max\{0, t\}$. We leave this equivalence as an exercise for the reader. This representation reveals the close link to the total variation distance, for which

$$\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} = \int \left|1 - \frac{d\mathbb{Q}}{d\mathbb{P}}\right| d\mathbb{P}(x) = 2 \int \left|1 - \frac{d\mathbb{Q}}{d\mathbb{P}}\right|_+ d\mathbb{P}(x).$$

An especially interesting aspect of the asymmetric coupling distance is that it satisfies a Pinsker-type inequality for product distributions. In particular, given any product distribution $\mathbb{P}$ in n variables, we have

$$\max\{C(\mathbb{Q}, \mathbb{P}), C(\mathbb{P}, \mathbb{Q})\} \leq \sqrt{2D(\mathbb{Q} \,\|\, \mathbb{P})} \tag{3.73}$$

for all distributions $\mathbb{Q}$ in n dimensions. This deep result is due to Samson; see the bibliographic section for further discussion. While simple to state, it is non-trivial to prove, and has some very powerful consequences for the concentration of convex and Lipschitz functions, as summarized in the following:

Theorem 3.24 *Consider a vector of independent random variables* $(X_1, \ldots, X_n)$*, each taking values in* $[0, 1]$*, and let* $f\colon \mathbb{R}^n \to \mathbb{R}$ *be convex, and L-Lipschitz with respect to the Euclidean norm. Then for all* $t \geq 0$*, we have*

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2e^{-\frac{t^2}{2L^2}}. \tag{3.74}$$

Remarks: Note that this is the analog of Theorem 2.26—namely, a dimension-independent

https://doi.org/10.1017/9781108627771.003 Published online by Cambridge University Press

form of concentration for Lipschitz functions of independent Gaussian variables, but formulated for *Lipschitz and convex* functions of bounded random variables.

Of course, the same bound also applies to a concave and Lipschitz function. Earlier, we saw that upper tail bounds can obtained under a slightly milder condition, namely that of separate convexity (see Theorem 3.4). However, two-sided tail bounds (or concentration inequalities) require these stronger convexity or concavity conditions, as imposed here.

Example 3.25 (Rademacher revisited) As previously introduced in Example 3.5, the Rademacher complexity of a set $\mathcal{A} \subseteq \mathbb{R}^n$ is defined in terms of the random variable

$$Z \equiv Z(\varepsilon_1, \ldots, \varepsilon_n) := \sup_{a \in \mathcal{A}} \sum_{k=1}^{n} a_k \varepsilon_k,$$

where $\{\varepsilon_k\}_{k=1}^n$ is an i.i.d. sequence of Rademacher variables. As shown in Example 3.5, the function $(\varepsilon_1, \ldots, \varepsilon_n) \mapsto Z(\varepsilon_1, \ldots, \varepsilon_n)$ is jointly convex, and Lipschitz with respect to the Euclidean norm with parameter $\mathcal{W}(\mathcal{A}) := \sup_{a \in \mathcal{A}} \|a\|_2$. Consequently, Theorem 3.24 implies that

$$\mathbb{P}[|Z - \mathbb{E}[Z]| \geq t] \leq 2 \exp\left(-\frac{t^2}{2\,\mathcal{W}^2(\mathcal{A})}\right). \tag{3.75}$$

Note that this bound sharpens our earlier inequality (3.17), both in terms of the exponent and in providing a two-sided result. ♣

Let us now prove Theorem 3.24.

Proof As defined, any Wasserstein distance immediately yields an upper bound on a quantity of the form $\int f(d\mathbb{Q} - d\mathbb{P})$, where f is a Lipschitz function. Although the asymmetric coupling-based distance is not a Wasserstein distance, the key fact is that it can be used to upper bound such differences when $f \colon [0,1]^n \to \mathbb{R}$ is Lipschitz and convex. Indeed, for a convex f, we have the lower bound $f(x) \geq f(y) + \langle \nabla f(y),\, x - y \rangle$, which implies that

$$f(y) - f(x) \leq \sum_{j=1}^{n} \left| \frac{\partial f}{\partial y_j}(y) \right| \mathbb{I}[x_j \neq y_j].$$

Here we have also used the fact that $|x_j - y_j| \leq \mathbb{I}[x_j \neq y_j]$ for variables taking values in the unit interval $[0,1]$. Consequently, for any coupling $\mathbb{M}$ of the pair $(\mathbb{P}, \mathbb{Q})$, we have

$$\begin{aligned}
\int f(y)\, d\mathbb{Q}(y) - \int f(x)\, d\mathbb{P}(x) &\leq \sum_{j=1}^{n} \left| \frac{\partial f}{\partial y_j}(y) \right| \mathbb{I}[x_j \neq y_j]\, d\mathbb{M}(x,y) \\
&= \int \sum_{j=1}^{n} \left| \frac{\partial f}{\partial y_j}(y) \right| \mathbb{M}[X_j \neq y_j \mid Y_j = y_j]\, d\mathbb{Q}(y) \\
&\leq \int \|\nabla f(y)\|_2 \sqrt{\sum_{j=1}^{n} \mathbb{M}^2[X_j \neq y_j \mid Y_j = y_j]}\, d\mathbb{Q}(y),
\end{aligned}$$

where we have applied the Cauchy–Schwarz inequality. By the Lipschitz condition and con-

https://doi.org/10.1017/9781108627771.003 Published online by Cambridge University Press

vexity, we have $\|\nabla f(y)\|_2 \leq L$ almost everywhere, and hence

$$\begin{aligned}\int f(y)\, d\mathbb{Q}(y) - \int f(x)\, d\mathbb{P}(x) &\leq L \int \left\{ \sum_{j=1}^{n} \mathbb{M}^2[X_j \neq y_j \mid Y_j = y_j] \right\}^{1/2} d\mathbb{Q}(y) \\ &\leq L \left[\int \sum_{j=1}^{n} \mathbb{M}^2[X_j \neq y_j \mid Y_j = y_j]\, d\mathbb{Q}(y) \right]^{1/2} \\ &= L C(\mathbb{P}, \mathbb{Q}).\end{aligned}$$

Consequently, the upper tail bound follows by a combination of the information inequality (3.73) and Theorem 3.19.

To obtain the lower bound for a convex Lipschitz function, it suffices to establish an upper bound for a concave Lipschitz function, say $g\colon [0,1]^n \to \mathbb{R}$. In this case, we have the upper bound

$$g(y) \leq g(x) + \langle \nabla g(x),\, y - x \rangle \leq g(x) + \sum_{j=1}^{n} \left| \frac{\partial g(x)}{\partial x_j} \right| \mathbb{I}[x_j \neq y_j],$$

and consequently

$$\int g\, d\mathbb{Q}(y) - \int g\, d\mathbb{P}(x) \leq \sum_{j=1}^{n} \left| \frac{\partial g(x)}{\partial x_j} \right| \mathbb{I}[x_j \neq y_j]\, d\mathbb{M}(x, y).$$

The same line of reasoning then shows that $\int g\, d\mathbb{Q}(y) - \int g\, d\mathbb{P}(x) \leq L C(\mathbb{Q}, \mathbb{P})$, from which the claim then follows as before. □

We have stated Theorem 3.24 for the familiar case of independent random variables. However, a version of the underlying information inequality (3.73) holds for many collections of random variables. In particular, consider an n-dimensional distribution $\mathbb{P}$ for which there exists some $\gamma > 0$ such that the following inequality holds:

$$\max\{C(\mathbb{Q}, \mathbb{P}),\, C(\mathbb{P}, \mathbb{Q})\} \leq \sqrt{2\gamma D(\mathbb{Q} \,\|\, \mathbb{P})} \qquad \text{for all distributions } \mathbb{Q}. \tag{3.76}$$

The same proof then shows that any L-Lipschitz function satisfies the concentration inequality

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2 \exp\left(-\frac{t^2}{2\gamma L^2} \right). \tag{3.77}$$

For example, for a Markov chain that satisfies the β-contraction condition (3.67), it can be shown that the information inequality (3.76) holds with $\gamma = (\frac{1}{1-\sqrt{\beta}})^2$. Consequently, any L-Lipschitz function (with respect to the Euclidean norm) of a β-contractive Markov chain satisfies the concentration inequality

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2 \exp\left(-\frac{(1 - \sqrt{\beta})^2 t^2}{2L^2} \right). \tag{3.78}$$

This bound is a dimension-independent analog of our earlier bound (3.69) for a contractive Markov chain. We refer the reader to the bibliographic section for further discussion of results of this type.

https://doi.org/10.1017/9781108627771.003 Published online by Cambridge University Press

3.4 Tail bounds for empirical processes

In this section, we illustrate the use of concentration inequalities in application to empirical processes. We encourage the interested reader to look ahead to Chapter 4 so as to acquire the statistical motivation for the classes of problems studied in this section. Here we use the entropy method to derive various tail bounds on the suprema of empirical processes—in particular, for random variables that are generated by taking suprema of sample averages over function classes. More precisely let $\mathscr{F}$ be a class of functions (each of the form $f\colon \mathcal{X} \to \mathbb{R}$), and let $(X_1, \ldots, X_n)$ be drawn from a product distribution $\mathbb{P} = \bigotimes_{i=1}^n \mathbb{P}_i$, where each $\mathbb{P}_i$ is supported on some set $\mathcal{X}_i \subseteq \mathcal{X}$. We then consider the random variable[5]

$$Z = \sup_{f \in \mathscr{F}} \left\{ \frac{1}{n} \sum_{i=1}^n f(X_i) \right\}. \tag{3.79}$$

The primary goal of this section is to derive a number of upper bounds on the tail event $\{Z \geq \mathbb{E}[Z] + \delta\}$.

As a passing remark, we note that, if the goal is to obtain bounds on the random variable $\sup_{f \in \mathscr{F}} \left|\frac{1}{n} \sum_{i=1}^n f(X_i)\right|$, then it can be reduced to an instance of the variable (3.79) by considering the augmented function class $\widetilde{\mathscr{F}} = \mathscr{F} \cup \{-\mathscr{F}\}$.

3.4.1 A functional Hoeffding inequality

We begin with the simplest type of tail bound for the random variable Z, namely one of the Hoeffding type. The following result is a generalization of the classical Hoeffding theorem for sums of bounded random variables.

Theorem 3.26 (Functional Hoeffding theorem) *For each $f \in \mathscr{F}$ and $i = 1, \ldots, n$, assume that there are real numbers $a_{i,f} \leq b_{i,f}$ such that $f(x) \in [a_{i,f}, b_{i,f}]$ for all $x \in \mathcal{X}_i$. Then for all $\delta \geq 0$, we have*

$$\mathbb{P}[Z \geq \mathbb{E}[Z] + \delta] \leq \exp\left(-\frac{n\delta^2}{4L^2}\right), \tag{3.80}$$

where $L^2 := \sup_{f \in \mathscr{F}} \{\frac{1}{n} \sum_{i=1}^n (b_{i,f} - a_{i,f})^2\}$.

Remark: In a very special case, Theorem 3.26 can be used to recover the classical Hoeffding inequality in the case of bounded random variables, albeit with a slightly worse constant. Indeed, if we let $\mathscr{F}$ be a singleton consisting of the identity function $f(x) = x$, then we have $Z = \frac{1}{n} \sum_{i=1}^n X_i$. Consequently, as long as $x_i \in [a_i, b_i]$, Theorem 3.26 implies that

$$\mathbb{P}\left[\frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \geq \delta\right] \leq e^{-\frac{n\delta^2}{4L^2}},$$

[5] Note that there can be measurability problems associated with this definition if $\mathscr{F}$ is not countable. See the bibliographic discussion in Chapter 4 for more details on how to resolve them.

https://doi.org/10.1017/9781108627771.003 Published online by Cambridge University Press

where $L^2 = \frac{1}{n}\sum_{i=1}^n (b_i - a_i)^2$. We thus recover the classical Hoeffding theorem, although the constant $1/4$ in the exponent is not optimal.

More substantive implications of Theorem 3.26 arise when it is applied to a larger function class $\mathscr{F}$. In order to appreciate its power, let us compare the upper tail bound (3.80) to the corresponding bound that can be derived from the bounded differences inequality, as applied to the function $(x_1, \ldots, x_n) \mapsto Z(x_1, \ldots, x_n)$. With some calculation, it can be seen that this function satisfies the bounded difference inequality with constant $L_i := \sup_{f\in\mathscr{F}} |b_{i,f} - a_{i,f}|$ in coordinate i. Consequently, the bounded differences method (Corollary 2.21) yields a sub-Gaussian tail bound, analogous to the bound (3.80), but with the parameter

$$\widetilde{L}^2 = \frac{1}{n}\sum_{i=1}^n \sup_{f\in\mathscr{F}} (b_{i,f} - a_{i,f})^2.$$

Note that the quantity $\widetilde{L}$—since it is defined by applying the supremum separately to each coordinate—can be substantially larger than the constant L defined in the theorem statement.

Proof It suffices to prove the result for a finite class of functions $\mathscr{F}$; the general result can be recovered by taking limits over an increasing sequence of such finite classes. Let us view Z as a function of the random variables $(X_1, \ldots, X_n)$. For each index $j = 1, \ldots, n$, define the random function

$$x_j \mapsto Z_j(x_j) = Z(X_1, \ldots, X_{j-1}, x_j, X_{j+1}, \ldots, X_n).$$

In order to avoid notational clutter, we work throughout this proof with the *unrescaled* version of Z, namely $Z = \sup_{f\in\mathscr{F}} \sum_{i=1}^n f(X_i)$. Combining the tensorization Lemma 3.8 with the bound (3.20a) from Lemma 3.7, we obtain

$$\mathbb{H}(e^{\lambda Z(X)}) \leq \lambda^2 \mathbb{E}\Bigg[\sum_{j=1}^n \mathbb{E}[(Z_j(X_j) - Z_j(Y_j))^2\, \mathbb{I}[Z_j(X_j) \geq Z_j(Y_j)]\, e^{\lambda Z(X)} \mid X^{\backslash j}]\Bigg]. \tag{3.81}$$

For each $f \in \mathscr{F}$, define the set $\mathscr{A}(f) := \{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid Z = \sum_{i=1}^n f(x_i)\}$, corresponding to the set of realizations for which the maximum defining Z is achieved by f. (If there are ties, then we resolve them arbitrarily so as to make the sets $\mathscr{A}(f)$ disjoint.) For any $x \in \mathscr{A}(f)$, we have

$$Z_j(x_j) - Z_j(y_j) = f(x_j) + \sum_{i\neq j}^n f(x_i) - \max_{\widetilde{f}\in\mathscr{F}}\Bigg\{\widetilde{f}(y_j) + \sum_{i\neq j}^n \widetilde{f}(x_i)\Bigg\} \leq f(x_j) - f(y_j).$$

As long as $Z_j(x_j) \geq Z_j(y_j)$, this inequality still holds after squaring both sides. Considering all possible sets $\mathscr{A}(f)$, we arrive at the upper bound

$$(Z_j(x_j) - Z_j(y_j))^2\, \mathbb{I}[Z_j(x_j) \geq Z_j(y_j)] \leq \sum_{f\in\mathscr{F}} \mathbb{I}[x \in \mathscr{A}(f)](f(x_j) - f(y_j))^2. \tag{3.82}$$

https://doi.org/10.1017/9781108627771.003 Published online by Cambridge University Press

Since $(f(x_j) - f(y_j))^2 \leq (b_{j,f} - a_{j,f})^2$ by assumption, summing over the indices j yields

$$\begin{aligned}\sum_{j=1}^{n} (Z_j(x_j) - Z_j(y_j))^2 \, \mathbb{I}[Z_k(x_k) \geq Z_k(y_k)] \, e^{\lambda Z(x)} &\leq \sum_{h \in \mathscr{F}} \mathbb{I}[x \in \mathcal{A}(h)] \sum_{k=1}^{n} (b_{k,h} - a_{k,h})^2 e^{\lambda Z(x)} \\ &\leq \sup_{f \in \mathscr{F}} \sum_{j=1}^{n} (b_{j,f} - a_{j,f})^2 e^{\lambda Z(x)} \\ &= nL^2 e^{\lambda Z(x)}.\end{aligned}$$

Substituting back into our earlier inequality (3.81), we find that

$$\mathbb{H}(e^{\lambda Z(X)}) \leq nL^2 \lambda^2 \, \mathbb{E}[e^{\lambda Z(X)}].$$

This is a sub-Gaussian entropy bound (3.5) with $\sigma = \sqrt{2n}\,L$, so that Proposition 3.2 implies that the unrescaled version of Z satisfies the tail bound

$$\mathbb{P}[Z \geq \mathbb{E}[Z] + t] \leq e^{-\frac{t^2}{4nL^2}}.$$

Setting $t = n\delta$ yields the claim (3.80) for the rescaled version of Z. □

3.4.2 A functional Bernstein inequality

In this section, we turn to the Bernstein refinement of the functional Hoeffding inequality from Theorem 3.26. As opposed to control only in terms of bounds on the function values, it also brings a notion of variance into play. As will be discussed at length in later chapters, this type of variance control plays a key role in obtaining sharp bounds for various types of statistical estimators.

Theorem 3.27 (Talagrand concentration for empirical processes) *Consider a countable class of functions $\mathscr{F}$ uniformly bounded by b. Then for all $\delta > 0$, the random variable* (3.79) *satisfies the upper tail bound*

$$\mathbb{P}[Z \geq \mathbb{E}[Z] + \delta] \leq 2 \exp\left(\frac{-n\delta^2}{8e\,\mathbb{E}[\Sigma^2] + 4b\delta} \right), \tag{3.83}$$

where $\Sigma^2 = \sup_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^{n} f^2(X_i)$.

In order to obtain a simpler bound, the expectation $\mathbb{E}[\Sigma^2]$ can be upper bounded. Using symmetrization techniques to be developed in Chapter 4, it can be shown that

$$\mathbb{E}[\Sigma^2] \leq \sigma^2 + 2b\,\mathbb{E}[Z], \tag{3.84}$$

where $\sigma^2 = \sup_{f \in \mathscr{F}} \mathbb{E}[f^2(X)]$. Using this upper bound on $\mathbb{E}[\Sigma^2]$ and performing some algebra, we obtain that there are universal positive constants (c_0, c_1) such that

$$\mathbb{P}[Z \geq \mathbb{E}[Z] + c_0\gamma\sqrt{t} + c_1 bt] \leq e^{-nt} \qquad \text{for all } t > 0, \tag{3.85}$$

https://doi.org/10.1017/9781108627771.003 Published online by Cambridge University Press

where $\gamma^2 = \sigma^2 + 2b\,\mathbb{E}[Z]$. See Exercise 3.16 for the derivation of this inequality from Theorem 3.27 and the upper bound (3.84). Although the proof outlined here leads to poor constants, the best known are $c_0 = \sqrt{2}$ and $c_1 = 1/3$; see the bibliographic section for further details.

In certain settings, it can be useful to exploit the bound (3.85) in an alternative form: in particular, for any $\epsilon > 0$, it implies the upper bound

$$\mathbb{P}[Z \geq (1+\epsilon)\mathbb{E}[Z] + c_0\sigma\sqrt{t} + (c_1 + c_0^2/\epsilon)bt] \leq e^{-nt}. \tag{3.86}$$

Conversely, we can recover the tail bound (3.85) by optimizing over $\epsilon > 0$ in the family of bounds (3.86); see Exercise 3.16 for the details of this equivalence.

Proof We assume without loss of generality that $b = 1$, since the general case can be reduced to this one. Moreover, as in the proof of Theorem 3.26, we work with the unrescaled version—namely, the variable $Z = \sup_{f\in\mathscr{F}} \sum_{i=1}^n f(X_i)$—and then translate our results back. Recall the definition of the sets $\mathcal{A}(f)$, and the upper bound (3.82) from the previous proof; substituting it into the entropy bound (3.81) yields the upper bound

$$\mathbb{H}(e^{\lambda Z}) \leq \lambda^2\,\mathbb{E}\Bigg[\sum_{j=1}^n \mathbb{E}\Bigg[\sum_{f\in\mathscr{F}} \mathbb{I}[x \in \mathcal{A}(f)](f(X_j) - f(Y_j))^2 e^{\lambda Z} \mid X^{\setminus j}\Bigg]\Bigg].$$

Now we have

$$\begin{aligned}\sum_{i=1}^n \sum_{f\in\mathscr{F}} \mathbb{I}[X \in \mathcal{A}(f)](f(X_j) - f(Y_j))^2 &\leq 2\sup_{f\in\mathscr{F}} \sum_{i=1}^n f^2(X_i) + 2\sup_{f\in\mathscr{F}} \sum_{i=1}^n f^2(Y_i)\\ &= 2\{\Gamma(X) + \Gamma(Y)\},\end{aligned}$$

where $\Gamma(X) := \sup_{f\in\mathscr{F}} \sum_{i=1}^n f^2(X_i)$ is the unrescaled version of Σ^2. Combined with our earlier inequality, we see that the entropy satisfies the upper bound

$$\mathbb{H}(e^{\lambda Z}) \leq 2\lambda^2\{\mathbb{E}[\Gamma e^{\lambda Z}] + \mathbb{E}[\Gamma]\,\mathbb{E}[e^{\lambda Z}]\}. \tag{3.87}$$

From the result of Exercise 3.4, we have $\mathbb{H}(e^{\lambda(Z+c)}) = e^{\lambda c}\mathbb{H}(e^{\lambda Z})$ for any constant $c \in \mathbb{R}$. Since the right-hand side also contains a term $e^{\lambda Z}$ in each component, we see that the same upper bound holds for $\mathbb{H}(e^{\lambda\widetilde{Z}})$, where $\widetilde{Z} = Z - \mathbb{E}[Z]$ is the centered version. We now introduce a lemma to control the term $\mathbb{E}[\Gamma e^{\lambda\widetilde{Z}}]$.

Lemma 3.28 (Controlling the random variance) *For all* $\lambda > 0$, *we have*

$$\mathbb{E}[\Gamma e^{\lambda\widetilde{Z}}] \leq (e-1)\mathbb{E}[\Gamma]\mathbb{E}[e^{\lambda\widetilde{Z}}] + \mathbb{E}[\widetilde{Z}e^{\lambda\widetilde{Z}}]. \tag{3.88}$$

Combining the upper bound (3.88) with the entropy upper bound (3.87) for $\widetilde{Z}$, we obtain

$$\mathbb{H}(e^{\lambda\widetilde{Z}}) \leq \lambda^2\{2e\,\mathbb{E}[\Gamma]\varphi(\lambda) + 2\varphi'(\lambda)\} \qquad \text{for all } \lambda > 0,$$

where $\varphi(\lambda) := \mathbb{E}[e^{\lambda\widetilde{Z}}]$ is the moment generating function of $\widetilde{Z}$. Since $\mathbb{E}[\widetilde{Z}] = 0$, we recognize this as an entropy bound of the Bernstein form (3.10) with $b = 2$ and $\sigma^2 = 2e\,\mathbb{E}[\Gamma]$.

Consequently, by the consequence (3.12) stated following Proposition 3.3, we conclude that

$$\mathbb{P}[\widetilde{Z} \geq \mathbb{E}[\widetilde{Z}] + \delta] \leq \exp\left(-\frac{\delta^2}{8e\,\mathbb{E}[\Gamma] + 4\delta}\right) \qquad \text{for all } \delta \geq 0.$$

Recalling the definition of Γ and rescaling by $1/n$, we obtain the stated claim of the theorem with $b = 1$.

It remains to prove Lemma 3.28. Consider the function $g(t) = e^t$ with conjugate dual $g^*(s) = s \log s - s$ for $s > 0$. By the definition of conjugate duality (also known as Young's inequality), we have $st \leq s \log s - s + e^t$ for all $s > 0$ and $t \in \mathbb{R}$. Applying this inequality with $s = e^{\lambda \widetilde{Z}}$ and $t = \Gamma - (e-1)\mathbb{E}[\Gamma]$ and then taking expectations, we find that

$$\mathbb{E}[\Gamma e^{\lambda \widetilde{Z}}] - (e-1)\mathbb{E}[e^{\lambda \widetilde{Z}}]\,\mathbb{E}[\Gamma] \leq \lambda \mathbb{E}[\widetilde{Z} e^{\lambda \widetilde{Z}}] - \mathbb{E}[e^{\lambda \widetilde{Z}}] + \mathbb{E}[e^{\Gamma - (e-1)\mathbb{E}[\Gamma]}].$$

Note that Γ is defined as a supremum of a class of functions taking values in $[0, 1]$. Therefore, by the result of Exercise 3.15, we have $\mathbb{E}[e^{\Gamma - (e-1)\mathbb{E}[\Gamma]}] \leq 1$. Moreover, by Jensen's inequality, we have $\mathbb{E}[e^{\lambda \widetilde{Z}}] \geq e^{\lambda \mathbb{E}[\widetilde{Z}]} = 1$. Putting together the pieces yields the claim (3.88). □

3.5 Bibliographic details and background

Concentration of measure is an extremely rich and deep area with an extensive literature; we refer the reader to the books by Ledoux (2001) and Boucheron et al. (2013) for more comprehensive treatments. Logarithmic Sobolev inequalities were introduced by Gross (1975) in a functional-analytic context. Their dimension-free nature makes them especially well suited for controlling infinite-dimensional stochastic processes (e.g., Holley and Stroock, 1987). The argument underlying the proof of Proposition 3.2 is based on the unpublished notes of Herbst. Ledoux (1996; 2001) pioneered the entropy method in application to a wider range of problems. The proof of Theorem 3.4 is based on Ledoux (1996), whereas the proofs of Lemmas 3.7 and 3.8 follow the book (Ledoux, 2001). A result of the form in Theorem 3.4 was initially proved by Talagrand (1991; 1995; 1996b) using his convex distance inequalities.

The Brunn–Minkowski theorem is a classical result from geometry and real analysis; see Gardner (2002) for a survey of its history and connections. Theorem 3.15 was proved independently by Prékopa (1971; 1973) and Leindler (1972). Brascamp and Lieb (1976) developed various connections between log-concavity and log-Sobolev inequalities; see the paper by Bobkov (1999) for further discussion. The inf-convolution argument underlying the proof of Theorem 3.16 was initiated by Maurey (1991), and further developed by Bobkov and Ledoux (2000). The lecture notes by Ball (1997) contain a wealth of information on geometric aspects of concentration, including spherical sections of convex bodies. Harper's theorem quoted in Example 3.13 is proven in the paper (Harper, 1966); it is a special case of a more general class of results known as discrete isoperimetric inequalities.

The Kantorovich–Rubinstein duality (3.55) was established by Kantorovich and Rubinstein (1958); it is a special case of more general results in optimal transport theory (e.g., Villani, 2008; Rachev and Ruschendorf, 1998). Marton (1996a) pioneered the use of the transportation cost method for deriving concentration inequalities, with subsequent contributions from various researchers (e.g., Dembo and Zeitouni, 1996; Dembo, 1997; Bobkov and Götze, 1999; Ledoux, 2001). See Marton's paper (1996b) for a proof of Theorem 3.22.

https://doi.org/10.1017/9781108627771.003 Published online by Cambridge University Press

The information inequality (3.73) was proved by Samson (2000). As noted following the statement of Theorem 3.24, he actually proves a much more general result, applicable to various types of dependent random variables. Other results on concentration for dependent random variables include the papers (Marton, 2004; Kontorovich and Ramanan, 2008).

Upper tail bounds on the suprema of empirical processes can be proved using chaining methods; see Chapter 5 for more details. Talagrand (1996a) initiated the use of concentration techniques to control deviations above the mean, as in Theorems 3.26 and 3.27. The theorems and entropy-based arguments given here are based on Chapter 7 of Ledoux (2001); the sketch in Exercise 3.15 is adapted from arguments in the same chapter. Sharper forms of Theorem 3.27 have been established by various authors (e.g., Massart, 2000; Bousquet, 2002, 2003; Klein and Rio, 2005). In particular, Bousquet (2003) proved that the bound (3.85) holds with constants $c_0 = \sqrt{2}$ and $c_1 = 1/3$. There are also various results on concentration of empirical processes for unbounded and/or dependent random variables (e.g., Adamczak, 2008; Mendelson, 2010); see also Chapter 14 for some one-sided results in this direction.

3.6 Exercises

Exercise 3.1 (Shannon entropy and Kullback–Leibler divergence) Given a discrete random variable $X \in \mathcal{X}$ with probability mass function p, its Shannon entropy is given by $H(X) := -\sum_{x\in\mathcal{X}} p(x)\log p(x)$. In this exercise, we explore the connection between the entropy functional $\mathbb{H}$ based on $\phi(u) = u\log u$ (see equation (3.2)) and the Shannon entropy.

(a) Consider the random variable $Z = p(U)$, where U is uniformly distributed over $\mathcal{X}$. Show that

$$\mathbb{H}(Z) = \frac{1}{|\mathcal{X}|}\{\log|\mathcal{X}| - H(X)\}.$$

(b) Use part (a) to show that Shannon entropy for a discrete random variable is maximized by a uniform distribution.

(c) Given two probability mass functions p and q, specify a choice of random variable Y such that $\mathbb{H}(Y) = D(p\,\|\,q)$, corresponding to the Kullback–Leibler divergence between p and q.

Exercise 3.2 (Chain rule and Kullback–Leibler divergence) Given two n-variate distributions $\mathbb{Q}$ and $\mathbb{P}$, show that the Kullback–Leibler divergence can be decomposed as

$$D(\mathbb{Q}\,\|\,\mathbb{P}) = D(\mathbb{Q}_1\,\|\,\mathbb{P}_1) + \sum_{j=2}^{n} \mathbb{E}_{\mathbb{Q}_1^{j-1}}[D(\mathbb{Q}_j(\cdot \mid X_1^{j-1})\,\|\,\mathbb{P}_j(\cdot \mid X_1^{j-1}))],$$

where $\mathbb{Q}_j(\cdot \mid X_1^{j-1})$ denotes the conditional distribution of X_j given $(X_1, \ldots, X_{j-1})$ under $\mathbb{Q}$, with a similar definition for $\mathbb{P}_j(\cdot \mid X_1^{j-1})$.

Exercise 3.3 (Variational representation for entropy) Show that the entropy has the variational representation

$$\mathbb{H}(e^{\lambda X}) = \inf_{t\in\mathbb{R}} \mathbb{E}[\psi(\lambda(X-t))e^{\lambda X}], \tag{3.89}$$

https://doi.org/10.1017/9781108627771.003 Published online by Cambridge University Press

where $\psi(u) := e^{-u} - 1 + u$.

Exercise 3.4 (Entropy and constant shifts) In this exercise, we explore some properties of the entropy.

(a) Show that for any random variable X and constant $c \in \mathbb{R}$,

$$\mathbb{H}(e^{\lambda(X+c)}) = e^{\lambda c}\, \mathbb{H}(e^{\lambda X}).$$

(b) Use part (a) to show that, if X satisfies the entropy bound (3.5), then so does $X + c$ for any constant c.

Exercise 3.5 (Equivalent forms of entropy) Let $\mathbb{H}_\varphi$ denote the entropy defined by the convex function $\varphi(u) = u \log u - u$. Show that $\mathbb{H}_\varphi(e^{\lambda X}) = \mathbb{H}(e^{\lambda X})$, where $\mathbb{H}$ denotes the usual entropy (defined by $\phi(u) = u \log u$).

Exercise 3.6 (Entropy rescaling) In this problem, we develop recentering and rescaling arguments used in the proof of Proposition 3.3.

(a) Show that a random variable X satisfies the Bernstein entropy bound (3.10) if and only if $\widetilde{X} = X - \mathbb{E}[X]$ satisfies the inequality

$$\mathbb{H}(e^{\lambda X}) \le \lambda^2\{b\varphi'_{\mathrm{x}}(\lambda) + \varphi_{\mathrm{x}}(\lambda)\sigma^2\} \qquad \text{for all } \lambda \in [0, 1/b). \tag{3.90}$$

(b) Show that a zero-mean random variable X satisfies inequality (3.90) if and only if $\widetilde{X} = X/b$ satisfies the bound

$$\mathbb{H}(e^{\lambda\widetilde{X}}) \le \lambda^2\{\varphi'_{\widetilde{X}}(\lambda) + \tilde{\sigma}^2\varphi_{\widetilde{X}}(\lambda)\} \qquad \text{for all } \lambda \in [0, 1),$$

where $\tilde{\sigma}^2 = \sigma^2/b^2$.

Exercise 3.7 (Entropy for bounded variables) Consider a zero-mean random variable X taking values in a finite interval $[a, b]$ almost surely. Show that its entropy satisfies the bound $\mathbb{H}(e^{\lambda X}) \le \frac{\lambda^2\sigma^2}{2}\varphi_{\mathrm{x}}(\lambda)$ with $\sigma := (b - a)/2$. (*Hint:* You may find the result of Exercise 3.3 useful.)

Exercise 3.8 (Exponential families and entropy) Consider a random variable $Y \in \mathcal{Y}$ with an exponential family distribution of the form

$$p_\theta(y) = h(y)e^{\langle\theta,\, T(y)\rangle - \Phi(\theta)},$$

where $T\colon \mathcal{Y} \to \mathbb{R}^d$ defines the vector of sufficient statistics, the function h is fixed, and the density p_θ is taken with respect to base measure μ. Assume that the log normalization term $\Phi(\theta) = \log \int_{\mathcal{Y}} \exp(\langle\theta, T(y)\rangle)h(y)\mu(dy)$ is finite for all $\theta \in \mathbb{R}^d$, and suppose moreover that ∇A is Lipschitz with parameter L, meaning that

$$\|\nabla\Phi(\theta) - \nabla\Phi(\theta')\|_2 \le L\|\theta - \theta'\|_2 \qquad \text{for all } \theta, \theta' \in \mathbb{R}^d. \tag{3.91}$$

https://doi.org/10.1017/9781108627771.003 Published online by Cambridge University Press

(a) For fixed unit-norm vector $v \in \mathbb{R}^d$, consider the random variable $X = \langle v, T(Y)\rangle$. Show that

$$\mathbb{H}(e^{\lambda X}) \leq L\lambda^2 \varphi_x(\lambda) \qquad \text{for all } \lambda \in \mathbb{R}.$$

Conclude that X is sub-Gaussian with parameter $\sqrt{2L}$.

(b) Apply part (a) to establish the sub-Gaussian property for:

(i) the univariate Gaussian distribution $Y \sim \mathcal{N}(\mu, \sigma^2)$ (*Hint:* Viewing σ^2 as fixed, write it as a one-dimensional exponential family.)

(ii) the Bernoulli variable $Y \in \{0, 1\}$ with $\theta = \frac{1}{2} \log \frac{\mathbb{P}[Y=1]}{\mathbb{P}[Y=0]}$.

Exercise 3.9 (Another variational representation) Prove the following variational representation:

$$\mathbb{H}(e^{\lambda f(X)}) = \sup_{g}\{\mathbb{E}[g(X)e^{\lambda f(X)}] \mid \mathbb{E}[e^{g(X)}] \leq 1\},$$

where the supremum ranges over all measurable functions. Exhibit a function g at which the supremum is obtained. (*Hint:* The result of Exercise 3.5 and the notion of conjugate duality could be useful.)

Exercise 3.10 (Brunn–Minkowski and classical isoperimetric inequality) In this exercise, we explore the connection between the Brunn–Minkowski (BM) inequality and the classical isoperimetric inequality.

(a) Show that the BM inequality (3.43) holds if and only if

$$\text{vol}(A + B)^{1/n} \geq \text{vol}(A)^{1/n} + \text{vol}(B)^{1/n} \tag{3.92}$$

for all convex bodies A and B.

(b) Show that the BM inequality (3.43) implies the “weaker” inequality (3.45).

(c) Conversely, show that inequality (3.45) also implies the original BM inequality (3.43). (*Hint:* From part (a), it suffices to prove the inequality (3.92) for bodies A and B with strictly positive volumes. Consider applying inequality (3.45) to the rescaled bodies $C := \frac{A}{\text{vol}(A)}$ and $D := \frac{B}{\text{vol}(B)}$, and a suitable choice of λ.)

Exercise 3.11 (Concentration on the Euclidean ball) Consider the uniform measure $\mathbb{P}$ over the Euclidean unit ball $\mathbb{B}_2^n = \{x \in \mathbb{R}^n \mid \|x\|_2 \leq 1\}$. In this example, we bound its concentration function using the Brunn–Minkowski inequality (3.45).

(a) Given any subset $A \subseteq \mathbb{B}_2^n$, show that

$$\frac{1}{2}\|a + b\|_2 \leq 1 - \frac{\epsilon^2}{8} \qquad \text{for all } a \in A \text{ and } b \in (A^\epsilon)^c.$$

To be clear, here we define $(A^\epsilon)^c := \mathbb{B}_2^n \backslash A^\epsilon$.

(b) Use the BM inequality (3.45) to show that $\mathbb{P}[A](1 - \mathbb{P}[A^\epsilon]) \leq (1 - \frac{\epsilon^2}{8})^{2n}$.

(c) Conclude that

$$\alpha_{\mathbb{P},(\mathcal{X},\rho)}(\epsilon) \leq 2e^{-n\epsilon^2/4} \qquad \text{for } \mathcal{X} = \mathbb{B}_2^n \text{ with } \rho(\cdot) = \|\cdot\|_2.$$

https://doi.org/10.1017/9781108627771.003 Published online by Cambridge University Press

Exercise 3.12 (Rademacher chaos variables) A symmetric positive semidefinite matrix $\mathbf{Q} \in \mathcal{S}_+^{d\times d}$ can be used to define a Rademacher chaos variable $X = \sum_{i,j=1}^d Q_{ij}\varepsilon_i\varepsilon_j$, where $\{\varepsilon_i\}_{i=1}^d$ are i.i.d. Rademacher variables.

(a) Prove that

$$\mathbb{P}[X \geq (\sqrt{\operatorname{trace}\mathbf{Q}} + t)^2] \leq 2\exp\left(-\frac{t^2}{16\,|||\mathbf{Q}|||_2}\right). \tag{3.93}$$

(b) Given an arbitrary symmetric matrix $\mathbf{M} \in \mathcal{S}^{d\times d}$, consider the decoupled Rademacher chaos variable $Y = \sum_{i=1,j=1}^d M_{ij}\varepsilon_i\varepsilon_j'$, where $\{\varepsilon_j'\}_{j=1}^d$ is a second i.i.d. Rademacher sequence, independent of the first. Show that

$$\mathbb{P}[Y \geq \delta] \leq 2\exp\left(-\frac{\delta^2}{4\,|||\mathbf{M}|||_{\mathrm{F}}^2 + 16\delta\,|||\mathbf{M}|||_2}\right).$$

(*Hint:* Part (a) could be useful in an intermediate step.)

Exercise 3.13 (Total variation and Wasserstein) Consider the Wasserstein distance based on the Hamming metric, namely $W_\rho(\mathbb{P}, \mathbb{Q}) = \inf_{\mathbb{M}} \mathbb{M}[X \neq Y]$, where the infimum is taken over all couplings $\mathbb{M}$—that is, distributions on the product space $\mathcal{X}\times\mathcal{X}$ with marginals $\mathbb{P}$ and $\mathbb{Q}$, respectively. Show that

$$\inf_{\mathbb{M}} \mathbb{M}[X \neq Y] = \|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} = \sup_A |\mathbb{P}(A) - \mathbb{Q}(A)|,$$

where the supremum ranges over all measurable subsets A of $\mathcal{X}$.

Exercise 3.14 (Alternative proof) In this exercise, we work through an alternative proof of Proposition 3.20. As noted, it suffices to consider the case $n = 2$. Let $\mathbb{P} = \mathbb{P}_1 \otimes \mathbb{P}_2$ be a product distribution, and let $\mathbb{Q}$ be an arbitrary distribution on $\mathcal{X}\times\mathcal{X}$.

(a) Show that the Wasserstein distance $W_\rho(\mathbb{Q}, \mathbb{P})$ is upper bounded by

$$\sup_{\|f\|_{\mathrm{Lip}}\leq 1}\left\{\int\left[\int f(x_1, x_2)\,(d\mathbb{Q}_{2|1} - d\mathbb{P}_2)\right]d\mathbb{Q}_1 + \int\left[\int f(x_1, x_2)\,d\mathbb{P}_2\right](d\mathbb{Q}_1 - d\mathbb{P}_1)\right\},$$

where the supremum ranges over all functions that are 1-Lipschitz with respect to the metric $\rho(x, x') = \sum_{i=1}^2 \rho_i(x_i, x_i')$.

(b) Use part (a) to show that

$$W_\rho(\mathbb{Q}, \mathbb{P}) \leq \left[\int \sqrt{2\gamma_2 D(\mathbb{Q}_{2|1}\,\|\,\mathbb{P}_2)}\,d\mathbb{Q}_1\right] + \sqrt{2\gamma_1 D(\mathbb{Q}_1\,\|\,\mathbb{P}_1)}.$$

(c) Complete the proof using part (b). (*Hint:* Cauchy–Schwarz and Exercise 3.2 could be useful.

Exercise 3.15 (Bounds for suprema of non-negative functions) Consider a random variable of the form $Z = \sup_{f\in\mathscr{F}} \sum_{i=1}^n f(V_i)$ where $\{V_i\}_{i=1}^n$ is an i.i.d. sequence of random variables, and $\mathscr{F}$ is a class of functions taking values in the interval $[0, 1]$. In this exercise, we prove that

$$\log \mathbb{E}[e^{\lambda Z}] \leq (e^\lambda - 1)\mathbb{E}[Z] \qquad \text{for any } \lambda \geq 0. \tag{3.94}$$

https://doi.org/10.1017/9781108627771.003 Published online by Cambridge University Press

As in our main development, we can reduce the problem to a finite class of functions $\mathscr{F}$, say with M functions $\{f^1, \ldots, f^M\}$. Defining the random vectors $X_i = (f^1(V_i), \ldots, f^M(V_i)) \in \mathbb{R}^M$ for $i = 1, \ldots, n$, we can then consider the function $Z(X) = \max_{j=1,\ldots,M} \sum_{i=1}^n X_i^j$. We let Z_k denote the function $X_k \mapsto Z(X)$ with all other X_i for $i \neq k$ fixed.

(a) Define $Y_k(X) := (X_1, \ldots, X_{k-1}, 0, X_{k+1}, X_n)$. Explain why $Z(X) - Z(Y_k(X)) \geq 0$.

(b) Use the tensorization approach and the variational representation from Exercise 3.3 to show that

$$\mathbb{H}(e^{\lambda Z(X)}) \leq \mathbb{E}\left[\sum_{k=1}^n \mathbb{E}[\psi(\lambda(Z(X) - Z(Y_k(X))))e^{\lambda Z(X)} \mid X^{\setminus k}]\right] \qquad \text{for all } \lambda > 0.$$

(c) For each $\ell = 1, \ldots, M$, let

$$\mathbb{A}_\ell = \left\{x = (x_1, \ldots, x_n) \in \mathbb{R}^{M \times n} \,\middle|\, \sum_{i=1}^n x_i^\ell = \max_{j=1,\ldots,M} \sum_{i=1}^n x_i^j\right\}.$$

Prove that

$$0 \leq \lambda\{Z(X) - Z(Y_k(X))\} \leq \lambda \sum_{\ell=1}^M \mathbb{I}[X \in \mathbb{A}_\ell] X_k^\ell \qquad \text{valid for all } \lambda \geq 0.$$

(d) Noting that $\psi(t) = e^{-t} + 1 - t$ is non-negative with $\psi(0) = 0$, argue by the convexity of ψ that

$$\psi(\lambda(Z(X) - Z(Y_k(X)))) \leq \psi(\lambda)\left[\sum_{\ell=1}^M \mathbb{I}[X \in \mathbb{A}_\ell] X_k^\ell\right] \qquad \text{for all } \lambda \geq 0.$$

(e) Combining with previous parts, prove that

$$\mathbb{H}(e^{\lambda Z}) \leq \psi(\lambda) \sum_{k=1}^n \mathbb{E}\left[\sum_{\ell=1}^M \mathbb{I}[X \in \mathbb{A}_\ell] X_k^\ell e^{\lambda Z(X)}\right] = \psi(\lambda)\mathbb{E}[Z(X)e^{\lambda Z(X)}].$$

(*Hint:* Observe that $\sum_{k=1}^n \sum_{\ell=1}^M \mathbb{I}[X \in \mathbb{A}_\ell] X_k^\ell = Z(X)$ by definition of the sets $\mathbb{A}_\ell$.)

(f) Use part (e) to show that $\varphi_Z(\lambda) = \mathbb{E}[e^{\lambda Z}]$ satisfies the differential inequality

$$[\log \varphi_Z(\lambda)]' \leq \frac{e^\lambda}{e^\lambda - 1} \log \varphi_Z(\lambda) \qquad \text{for all } \lambda > 0,$$

and use this to complete the proof.

Exercise 3.16 (Different forms of functional Bernstein) Consider a random variable Z that satisfies a Bernstein tail bound of the form

$$\mathbb{P}[Z \geq \mathbb{E}[Z] + \delta] \leq \exp\left(-\frac{n\delta^2}{c_1\gamma^2 + c_2 b\delta}\right) \qquad \text{for all } \delta \geq 0,$$

where c_1 and c_2 are universal constants.

(a) Show that

$$\mathbb{P}\left[Z \geq \mathbb{E}[Z] + \gamma\sqrt{\frac{c_1 t}{n}} + \frac{c_2 b t}{n}\right] \leq e^{-t} \qquad \text{for all } t \geq 0. \tag{3.95a}$$

https://doi.org/10.1017/9781108627771.003 Published online by Cambridge University Press

(b) If, in addition, $\gamma^2 \le \sigma^2 + c_3 b\mathbb{E}[Z]$, we have

$$\mathbb{P}\left[Z \ge (1+\epsilon)\mathbb{E}[Z] + \sigma\sqrt{\frac{c_1 t}{n}} + \left(c_2 + \frac{c_1 c_3}{2\epsilon}\right)\frac{bt}{n}\right] \le e^{-t} \quad \text{for all } t \ge 0 \text{ and } \epsilon > 0. \tag{3.95b}$$

https://doi.org/10.1017/9781108627771.003 Published online by Cambridge University Press