6

Random matrices and covariance estimation

Covariance matrices play a central role in statistics, and there exist a variety of methods for estimating them based on data. The problem of covariance estimation dovetails with random matrix theory, since the sample covariance is a particular type of random matrix. A classical framework allows the sample size n to tend to infinity while the matrix dimension d remains fixed; in such a setting, the behavior of the sample covariance matrix is characterized by the usual limit theory. By contrast, for high-dimensional random matrices in which the data dimension is either comparable to the sample size ($d \asymp n$), or possibly much larger than the sample size ($d \gg n$), many new phenomena arise.

High-dimensional random matrices play an important role in many branches of science, mathematics and engineering, and have been studied extensively. Part of high-dimensional theory is asymptotic in nature, such as the Wigner semicircle law and the Marčenko–Pastur law for the asymptotic distribution of the eigenvalues of a sample covariance matrix (see Chapter 1 for illustration of the latter). By contrast, this chapter is devoted to an exploration of random matrices in a non-asymptotic setting, with the goal of obtaining explicit deviation inequalities that hold for all sample sizes and matrix dimensions. Beginning with the simplest case—namely ensembles of Gaussian random matrices—we then discuss more general sub-Gaussian ensembles, and then move onwards to ensembles with milder tail conditions. Throughout our development, we bring to bear the techniques from concentration of measure, comparison inequalities and metric entropy developed previously in Chapters 2 through 5. In addition, this chapter introduces new some techniques, among them a class of matrix tail bounds developed over the past decade (see Section 6.4).

6.1 Some preliminaries

We begin by introducing notation and preliminary results used throughout this chapter, before setting up the problem of covariance estimation more precisely.

6.1.1 Notation and basic facts

Given a rectangular matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$ with $n \geq m$, we write its ordered singular values as

$$\sigma_{\max}(\mathbf{A}) = \sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq \cdots \geq \sigma_m(\mathbf{A}) = \sigma_{\min}(\mathbf{A}) \geq 0.$$

Note that the minimum and maximum singular values have the variational characterization

$$\sigma_{\max}(\mathbf{A}) = \max_{v\in\mathbb{S}^{m-1}} \|\mathbf{A}v\|_2 \quad \text{and} \quad \sigma_{\min}(\mathbf{A}) = \min_{v\in\mathbb{S}^{m-1}} \|\mathbf{A}v\|_2, \tag{6.1}$$

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press

where $\mathbb{S}^{d-1} := \{v \in \mathbb{R}^d \mid \|v\|_2 = 1\}$ is the Euclidean unit sphere in $\mathbb{R}^d$. Note that we have the equivalence $|||\mathbf{A}|||_2 = \sigma_{\max}(\mathbf{A})$.

Since covariance matrices are symmetric, we also focus on the set of symmetric matrices in $\mathbb{R}^d$, denoted $\mathcal{S}^{d\times d} := \{\mathbf{Q} \in \mathbb{R}^{d\times d} \mid \mathbf{Q} = \mathbf{Q}^{\mathrm{T}}\}$, as well as the subset of positive semidefinite matrices given by

$$\mathcal{S}_+^{d\times d} := \{\mathbf{Q} \in \mathcal{S}^{d\times d} \mid \mathbf{Q} \succeq 0\}. \tag{6.2}$$

From standard linear algebra, we recall the facts that any matrix $\mathbf{Q} \in \mathcal{S}^{d\times d}$ is diagonalizable via a unitary transformation, and we use $\gamma(\mathbf{Q}) \in \mathbb{R}^d$ to denote its vector of eigenvalues, ordered as

$$\gamma_{\max}(\mathbf{Q}) = \gamma_1(\mathbf{Q}) \geq \gamma_2(\mathbf{Q}) \geq \cdots \geq \gamma_d(\mathbf{Q}) = \gamma_{\min}(\mathbf{Q}).$$

Note that a matrix $\mathbf{Q}$ is positive semidefinite—written $\mathbf{Q} \succeq 0$ for short—if and only if $\gamma_{\min}(\mathbf{Q}) \geq 0$.

Our analysis frequently exploits the Rayleigh–Ritz variational characterization of the minimum and maximum eigenvalues—namely

$$\gamma_{\max}(\mathbf{Q}) = \max_{v\in\mathbb{S}^{d-1}} v^{\mathrm{T}}\mathbf{Q}v \quad \text{and} \quad \gamma_{\min}(\mathbf{Q}) = \min_{v\in\mathbb{S}^{d-1}} v^{\mathrm{T}}\mathbf{Q}v. \tag{6.3}$$

For any symmetric matrix $\mathbf{Q}$, the ℓ_2-operator norm can be written as

$$|||\mathbf{Q}|||_2 = \max\{\gamma_{\max}(\mathbf{Q}), |\gamma_{\min}(\mathbf{Q})|\}, \tag{6.4a}$$

by virtue of which it inherits the variational representation

$$|||\mathbf{Q}|||_2 := \max_{v\in\mathbb{S}^{d-1}} \left|v^{\mathrm{T}}\mathbf{Q}v\right|. \tag{6.4b}$$

Finally, given a rectangular matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$ with $n \geq m$, suppose that we define the m-dimensional symmetric matrix $\mathbf{R} := \mathbf{A}^{\mathrm{T}}\mathbf{A}$. We then have the relationship

$$\gamma_j(\mathbf{R}) = (\sigma_j(\mathbf{A}))^2 \qquad \text{for } j = 1, \ldots, m.$$

6.1.2 Set-up of covariance estimation

Let us now define the problem of covariance matrix estimation. Let $\{x_1, \ldots, x_n\}$ be a collection of n independent and identically distributed samples[1] from a distribution in $\mathbb{R}^d$ with zero mean, and covariance matrix $\mathbf{\Sigma} = \mathrm{cov}(x_1) \in \mathcal{S}_+^{d\times d}$. A standard estimator of $\mathbf{\Sigma}$ is the *sample covariance matrix*

$$\widehat{\mathbf{\Sigma}} := \frac{1}{n}\sum_{i=1}^{n} x_i x_i^{\mathrm{T}}. \tag{6.5}$$

Since each x_i has zero mean, we are guaranteed that $\mathbb{E}[x_i x_i^T] = \mathbf{\Sigma}$, and hence that the random matrix $\widehat{\mathbf{\Sigma}}$ is an unbiased estimator of the population covariance $\mathbf{\Sigma}$. Consequently, the error matrix $\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}$ has mean zero, and our goal in this chapter is to obtain bounds on the error

[1] In this chapter, we use a lower case x to denote a random vector, so as to distinguish it from a random matrix.

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press

measured in the ℓ_2-operator norm. By the variational representation (6.4b), a bound of the form $|||\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}|||_2 \leq \epsilon$ is equivalent to asserting that

$$\max_{v \in \mathbb{S}^{d-1}} \left| \frac{1}{n} \sum_{i=1}^{n} \langle x_i,\, v_i \rangle^2 - v^{\mathrm{T}} \mathbf{\Sigma} v \right| \leq \epsilon. \tag{6.6}$$

This representation shows that controlling the deviation $|||\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}|||_2$ is equivalent to establishing a uniform law of large numbers for the class of functions $x \mapsto \langle x,\, v\rangle^2$, indexed by vectors $v \in \mathbb{S}^{d-1}$. See Chapter 4 for further discussion of such uniform laws in a general setting.

Control in the operator norm also guarantees that the eigenvalues of $\widehat{\mathbf{\Sigma}}$ are uniformly close to those of $\mathbf{\Sigma}$. In particular, by a corollary of Weyl's theorem (see the bibliographic section for details), we have

$$\max_{j=1,\ldots,d} |\gamma_j(\widehat{\mathbf{\Sigma}}) - \gamma_j(\mathbf{\Sigma})| \leq |||\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}|||_2. \tag{6.7}$$

A similar type of guarantee can be made for the eigenvectors of the two matrices, but only if one has additional control on the separation between adjacent eigenvalues. See our discussion of principal component analysis in Chapter 8 for more details.

Finally, we point out the connection to the singular values of the random matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, denoted by $\{\sigma_j(\mathbf{X})\}_{j=1}^{\min\{n,d\}}$. Since the matrix $\mathbf{X}$ has the vector x_i^{T} as its ith row, we have

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^{\mathrm{T}} = \frac{1}{n} \mathbf{X}^{\mathrm{T}} \mathbf{X},$$

and hence it follows that the eigenvalues of $\widehat{\mathbf{\Sigma}}$ are the squares of the singular values of $\mathbf{X}/\sqrt{n}$.

6.2 Wishart matrices and their behavior

We begin by studying the behavior of singular values for random matrices with Gaussian rows. More precisely, let us suppose that each sample x_i is drawn i.i.d. from a multivariate $\mathcal{N}(0, \mathbf{\Sigma})$ distribution, in which case we say that the associated matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, with x_i^{T} as its ith row, is drawn from the $\mathbf{\Sigma}$-*Gaussian ensemble*. The associated sample covariance $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is said to follow a multivariate Wishart distribution.

Theorem 6.1 *Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be drawn according to the $\mathbf{\Sigma}$-Gaussian ensemble. Then for all $\delta > 0$, the maximum singular value $\sigma_{\max}(\mathbf{X})$ satisfies the upper deviation inequality*

$$\mathbb{P}\left[\frac{\sigma_{\max}(\mathbf{X})}{\sqrt{n}} \geq \gamma_{\max}(\sqrt{\mathbf{\Sigma}})\,(1 + \delta) + \sqrt{\frac{\mathrm{tr}(\mathbf{\Sigma})}{n}} \right] \leq e^{-n\delta^2/2}. \tag{6.8}$$

Moreover, for $n \geq d$, the minimum singular value $\sigma_{\min}(\mathbf{X})$ satisfies the analogous lower

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press

deviation inequality

$$\mathbb{P}\left[\frac{\sigma_{\min}(\mathbf{X})}{\sqrt{n}} \leq \gamma_{\min}(\sqrt{\mathbf{\Sigma}})\,(1-\delta) - \sqrt{\frac{\operatorname{tr}(\mathbf{\Sigma})}{n}}\right] \leq e^{-n\delta^2/2}. \tag{6.9}$$

Before proving this result, let us consider some illustrative examples.

Example 6.2 (Operator norm bounds for the standard Gaussian ensemble) Consider a random matrix $\mathbf{W} \in \mathbb{R}^{n\times d}$ generated with i.i.d. $\mathcal{N}(0,1)$ entries. This choice yields an instance of $\mathbf{\Sigma}$-Gaussian ensemble, in particular with $\mathbf{\Sigma} = \mathbf{I}_d$. By specializing Theorem 6.1, we conclude that for $n \geq d$, we have

$$\frac{\sigma_{\max}(\mathbf{W})}{\sqrt{n}} \leq 1 + \delta + \sqrt{\frac{d}{n}} \quad \text{and} \quad \frac{\sigma_{\min}(\mathbf{W})}{\sqrt{n}} \geq 1 - \delta - \sqrt{\frac{d}{n}}, \tag{6.10}$$

where both bounds hold with probability greater than $1 - 2e^{-n\delta^2/2}$. These bounds on the singular values of $\mathbf{W}$ imply that

$$\left|\!\left|\!\left| \frac{1}{n}\mathbf{W}^{\mathrm{T}}\mathbf{W} - \mathbf{I}_d \right|\!\right|\!\right|_2 \leq 2\epsilon + \epsilon^2, \qquad \text{where } \epsilon = \sqrt{\frac{d}{n}} + \delta, \tag{6.11}$$

with the same probability. Consequently, the sample covariance $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\mathbf{W}^{\mathrm{T}}\mathbf{W}$ is a consistent estimate of the identity matrix $\mathbf{I}_d$ whenever $d/n \to 0$. ♣

The preceding example has interesting consequences for the problem of sparse linear regression using standard Gaussian random matrices, as in compressed sensing; in particular, see our discussion of the restricted isometry property in Chapter 7. On the other hand, from the perspective of covariance estimation, estimating the identity matrix is not especially interesting. However, a minor modification does lead to a more realistic family of problems.

Example 6.3 (Gaussian covariance estimation) Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be a random matrix from the $\mathbf{\Sigma}$-Gaussian ensemble. By standard properties of the multivariate Gaussian, we can write $\mathbf{X} = \mathbf{W}\sqrt{\mathbf{\Sigma}}$, where $\mathbf{W} \in \mathbb{R}^{n\times d}$ is a standard Gaussian random matrix, and hence

$$\left|\!\left|\!\left| \frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X} - \mathbf{\Sigma} \right|\!\right|\!\right|_2 = \left|\!\left|\!\left| \sqrt{\mathbf{\Sigma}}\left(\frac{1}{n}\mathbf{W}^{\mathrm{T}}\mathbf{W} - \mathbf{I}_d\right)\sqrt{\mathbf{\Sigma}} \right|\!\right|\!\right|_2 \leq |\!|\!|\mathbf{\Sigma}|\!|\!|_2 \left|\!\left|\!\left| \frac{1}{n}\mathbf{W}^{\mathrm{T}}\mathbf{W} - \mathbf{I}_d \right|\!\right|\!\right|_2.$$

Consequently, by exploiting the bound (6.11), we are guaranteed that, for all $\delta > 0$,

$$\frac{|\!|\!|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}|\!|\!|_2}{|\!|\!|\mathbf{\Sigma}|\!|\!|_2} \leq 2\sqrt{\frac{d}{n}} + 2\delta + \left(\sqrt{\frac{d}{n}} + \delta\right)^2, \tag{6.12}$$

with probability at least $1 - 2e^{-n\delta^2/2}$. Overall, we conclude that the relative error $|\!|\!|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}|\!|\!|_2/|\!|\!|\mathbf{\Sigma}|\!|\!|_2$ converges to zero as long the ratio d/n converges to zero. ♣

It is interesting to consider Theorem 6.1 in application to sequences of matrices that satisfy additional structure, one being control on the eigenvalues of the covariance matrix $\mathbf{\Sigma}$.

Example 6.4 (Faster rates under trace constraints) Recall that $\{\gamma_j(\mathbf{\Sigma})\}_{j=1}^d$ denotes the ordered sequence of eigenvalues of the matrix $\mathbf{\Sigma}$, with $\gamma_1(\mathbf{\Sigma})$ being the maximum eigenvalue.

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press

Now consider a non-zero covariance matrix $\mathbf{\Sigma}$ that satisfies a "trace constraint" of the form

$$\frac{\operatorname{tr}(\mathbf{\Sigma})}{|||\mathbf{\Sigma}|||_2} = \frac{\sum_{j=1}^{d} \gamma_j(\mathbf{\Sigma})}{\gamma_1(\mathbf{\Sigma})} \le C, \tag{6.13}$$

where C is some constant independent of dimension. Note that this ratio is a rough measure of the matrix rank, since inequality (6.13) always holds with $C = \operatorname{rank}(\mathbf{\Sigma})$. Perhaps more interesting are matrices that are full-rank but that exhibit a relatively fast eigendecay, with a canonical instance being matrices that belong to the Schatten q-"balls" of matrices. For symmetric matrices, these sets take the form

$$\mathbb{B}_q(R_q) := \left\{ \mathbf{\Sigma} \in \mathcal{S}^{d\times d} \;\middle|\; \sum_{j=1}^{d} |\gamma_j(\mathbf{\Sigma})|^q \le R_q \right\}, \tag{6.14}$$

where $q \in [0, 1]$ is a given parameter, and $R_q > 0$ is the radius. If we restrict to matrices with eigenvalues in $[-1, 1]$, these matrix families are nested: the smallest set with $q = 0$ corresponds to the case of matrices with rank at most R_0, whereas the other extreme $q = 1$ corresponds to an explicit trace constraint. Note that any non-zero matrix $\mathbf{\Sigma} \in \mathbb{B}_q(R_q)$ satisfies a bound of the form (6.13) with the parameter $C = R_q/(\gamma_1(\mathbf{\Sigma}))^q$.

For any matrix class satisfying the bound (6.13), Theorem 6.1 guarantees that, with high probability, the maximum singular value is bounded above as

$$\frac{\sigma_{\max}(\mathbf{X})}{\sqrt{n}} \le \gamma_{\max}(\sqrt{\mathbf{\Sigma}}) \left(1 + \delta + \sqrt{\frac{C}{n}} \right). \tag{6.15}$$

By comparison to the earlier bound (6.10) for $\mathbf{\Sigma} = \mathbf{I}_d$, we conclude that the parameter C plays the role of the *effective dimension*. ♣

We now turn to the proof of Theorem 6.1.

Proof In order to simplify notation in the proof, let us introduce the convenient shorthand $\bar{\sigma}_{\max} = \gamma_{\max}(\sqrt{\mathbf{\Sigma}})$ and $\bar{\sigma}_{\min} = \gamma_{\min}(\sqrt{\mathbf{\Sigma}})$. Our proofs of both the upper and lower bounds consist of two steps: first, we use concentration inequalities (see Chapter 2) to argue that the random singular value is close to its expectation with high probability, and second, we use Gaussian comparison inequalities (see Chapter 5) to bound the expected values.

Maximum singular value: As noted previously, by standard properties of the multivariate Gaussian distribution, we can write $\mathbf{X} = \mathbf{W}\sqrt{\mathbf{\Sigma}}$, where the random matrix $\mathbf{W} \in \mathbb{R}^{n\times d}$ has i.i.d. $\mathcal{N}(0, 1)$ entries. Now let us view the mapping $\mathbf{W} \mapsto \frac{\sigma_{\max}(\mathbf{W}\sqrt{\mathbf{\Sigma}})}{\sqrt{n}}$ as a real-valued function on $\mathbb{R}^{nd}$. By the argument given in Example 2.32, this function is Lipschitz with respect to the Euclidean norm with constant at most $L = \bar{\sigma}_{\max}/\sqrt{n}$. By concentration of measure for Lipschitz functions of Gaussian random vectors (Theorem 2.26), we conclude that

$$\mathbb{P}[\sigma_{\max}(\mathbf{X}) \ge \mathbb{E}[\sigma_{\max}(\mathbf{X})] + \sqrt{n}\bar{\sigma}_{\max}\delta] \le e^{-n\delta^2/2}.$$

Consequently, it suffices to show that

$$\mathbb{E}[\sigma_{\max}(\mathbf{X})] \le \sqrt{n}\bar{\sigma}_{\max} + \sqrt{\operatorname{tr}(\mathbf{\Sigma})}. \tag{6.16}$$

In order to do so, we first write $\sigma_{\max}(\mathbf{X})$ in a variational fashion, as the maximum of a

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press

suitably defined Gaussian process. By definition of the maximum singular value, we have $\sigma_{\max}(\mathbf{X}) = \max_{v' \in \mathbb{S}^{d-1}} \|\mathbf{X}v'\|_2$, where $\mathbb{S}^{d-1}$ denotes the Euclidean unit sphere in $\mathbb{R}^d$. Recalling the representation $\mathbf{X} = \mathbf{W}\sqrt{\mathbf{\Sigma}}$ and making the substitution $v = \sqrt{\mathbf{\Sigma}}\, v'$, we can write

$$\sigma_{\max}(\mathbf{X}) = \max_{v \in \mathbb{S}^{d-1}(\mathbf{\Sigma}^{-1})} \|\mathbf{W}v\|_2 = \max_{u \in \mathbb{S}^{n-1}} \max_{v \in \mathbb{S}^{d-1}(\mathbf{\Sigma}^{-1})} \underbrace{u^{\mathrm{T}}\mathbf{W}v}_{Z_{u,v}},$$

where $\mathbb{S}^{d-1}(\mathbf{\Sigma}^{-1}) := \{v \in \mathbb{R}^d \mid \|\mathbf{\Sigma}^{-\frac{1}{2}}v\|_2 = 1\}$ is an ellipse. Consequently, obtaining bounds on the maximum singular value corresponds to controlling the supremum of the zero-mean Gaussian process $\{Z_{u,v}, (u,v) \in \mathbb{T}\}$ indexed by the set $\mathbb{T} := \mathbb{S}^{n-1} \times \mathbb{S}^{d-1}(\mathbf{\Sigma}^{-1})$.

We upper bound the expected value of this supremum by constructing another Gaussian process $\{Y_{u,v}, (u,v) \in \mathbb{T}\}$ such that $\mathbb{E}[(Z_{u,v} - Z_{\widetilde{u},\widetilde{v}})^2] \leq \mathbb{E}[(Y_{u,v} - Y_{\widetilde{u},\widetilde{v}})^2]$ for all pairs (u,v) and $(\widetilde{u},\widetilde{v})$ in $\mathbb{T}$. We can then apply the Sudakov–Fernique comparison (Theorem 5.27) to conclude that

$$\mathbb{E}[\sigma_{\max}(\mathbf{X})] = \mathbb{E}\left[\max_{(u,v)\in\mathbb{T}} Z_{u,v}\right] \leq \mathbb{E}\left[\max_{(u,v)\in\mathbb{T}} Y_{u,v}\right]. \tag{6.17}$$

Introducing the Gaussian process $Z_{u,v} := u^{\mathrm{T}}\mathbf{W}v$, let us first compute the induced pseudometric ρ_Z. Given two pairs (u,v) and $(\widetilde{u},\widetilde{v})$, we may assume without loss of generality that $\|v\|_2 \leq \|\widetilde{v}\|_2$. (If not, we simply reverse the roles of (u,v) and $(\widetilde{u},\widetilde{v})$ in the argument to follow.) We begin by observing that $Z_{u,v} = \langle\langle \mathbf{W},\ uv^{\mathrm{T}}\rangle\rangle$, where we use $\langle\langle A,\ B\rangle\rangle := \sum_{j=1}^{n}\sum_{k=1}^{d} A_{jk}B_{jk}$ to denote the trace inner product. Since the matrix $\mathbf{W}$ has i.i.d. $\mathcal{N}(0,1)$ entries, we have

$$\mathbb{E}[(Z_{u,v} - Z_{\widetilde{u},\widetilde{v}})^2] = \mathbb{E}[(\langle\langle \mathbf{W},\ uv^{\mathrm{T}} - \widetilde{u}\widetilde{v}^{\mathrm{T}}\rangle\rangle)^2] = |||uv^{\mathrm{T}} - \widetilde{u}\widetilde{v}^{\mathrm{T}}|||_F^2.$$

Rearranging and expanding out this Frobenius norm, we find that

$$\begin{aligned}
|||uv^{\mathrm{T}} - \widetilde{u}\widetilde{v}^{\mathrm{T}}|||_F^2 &= |||u(v-\widetilde{v})^{\mathrm{T}} + (u-\widetilde{u})\widetilde{v}^{\mathrm{T}}|||_{\mathrm{F}}^2 \\
&= |||(u-\widetilde{u})\widetilde{v}^{\mathrm{T}}|||_{\mathrm{F}}^2 + |||u(v-\widetilde{v})^{\mathrm{T}}|||_{\mathrm{F}}^2 + 2\langle\langle u(v-\widetilde{v})^{\mathrm{T}},\ (u-\widetilde{u})\widetilde{v}^{\mathrm{T}}\rangle\rangle \\
&\leq \|\widetilde{v}\|_2^2\,\|u-\widetilde{u}\|_2^2 + \|u\|_2^2\,\|v-\widetilde{v}\|_2^2 + 2(\|u\|_2^2 - \langle u,\ \widetilde{u}\rangle)(\langle v,\ \widetilde{v}\rangle - \|\widetilde{v}\|_2^2).
\end{aligned}$$

Now since $\|u\|_2 = \|\widetilde{u}\|_2 = 1$ by definition of the set $\mathbb{T}$, we have $\|u\|_2^2 - \langle u,\ \widetilde{u}\rangle \geq 0$. On the other hand, we have

$$|\langle v,\ \widetilde{v}\rangle| \overset{\text{(i)}}{\leq} \|v\|_2\,\|\widetilde{v}\|_2 \overset{\text{(ii)}}{\leq} \|\widetilde{v}\|_2^2,$$

where step (i) follows from the Cauchy–Schwarz inequality, and step (ii) follows from our initial assumption that $\|v\|_2 \leq \|\widetilde{v}\|_2$. Combined with our previous bound on $\|u\|_2^2 - \langle u,\ \widetilde{u}\rangle$, we conclude that

$$\underbrace{(\|u\|_2^2 - \langle u,\ \widetilde{u}\rangle)}_{\geq 0}\ \underbrace{(\langle v,\ \widetilde{v}\rangle - \|\widetilde{v}\|_2^2)}_{\leq 0} \leq 0.$$

Putting together the pieces, we conclude that

$$|||uv^{\mathrm{T}} - \widetilde{u}\widetilde{v}^{\mathrm{T}}|||_{\mathrm{F}}^2 \leq \|\widetilde{v}\|_2^2\,\|u-\widetilde{u}\|_2^2 + \|v-\widetilde{v}\|_2^2.$$

Finally, by definition of the set $\mathbb{S}^{d-1}(\mathbf{\Sigma}^{-1})$, we have $\|\widetilde{v}\|_2 \leq \bar{\sigma}_{\max} = \gamma_{\max}(\sqrt{\mathbf{\Sigma}})$, and hence

$$\mathbb{E}[(Z_{u,v} - Z_{\widetilde{u},\widetilde{v}})^2] \leq \bar{\sigma}_{\max}^2\|u-\widetilde{u}\|_2^2 + \|v-\widetilde{v}\|_2^2.$$

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press

Motivated by this inequality, we define the Gaussian process $Y_{u,v} := \bar{\sigma}_{\max} \langle g, u \rangle + \langle h, v \rangle$, where $g \in \mathbb{R}^n$ and $h \in \mathbb{R}^d$ are both standard Gaussian random vectors (i.e., with i.i.d. $\mathcal{N}(0, 1)$ entries), and mutually independent. By construction, we have

$$\mathbb{E}[(Y_\theta - Y_{\widetilde{\theta}})^2] = \bar{\sigma}^2_{\max} \|u - \widetilde{u}\|_2^2 + \|v - \widetilde{v}\|_2^2.$$

Thus, we may apply the Sudakov–Fernique bound (6.17) to conclude that

$$\begin{aligned}\mathbb{E}[\sigma_{\max}(\mathbf{X})] &\leq \mathbb{E}\Big[\sup_{(u,v)\in\mathbb{T}} Y_{u,v}\Big] \\ &= \bar{\sigma}_{\max}\, \mathbb{E}\Big[\sup_{u\in\mathbb{S}^{n-1}} \langle g, u\rangle\Big] + \mathbb{E}\Big[\sup_{v\in\mathbb{S}^{d-1}(\mathbf{\Sigma}^{-1})} \langle h, v\rangle\Big] \\ &= \bar{\sigma}_{\max} \mathbb{E}[\|g\|_2] + \mathbb{E}[\|\sqrt{\mathbf{\Sigma}} h\|_2]\end{aligned}$$

By Jensen's inequality, we have $\mathbb{E}[\|g\|_2] \leq \sqrt{n}$, and similarly,

$$\mathbb{E}[\|\sqrt{\mathbf{\Sigma}} h\|_2] \leq \sqrt{\mathbb{E}[h^{\mathrm{T}} \mathbf{\Sigma} h]} = \sqrt{\operatorname{tr}(\mathbf{\Sigma})},$$

which establishes the claim (6.16).

The lower bound on the minimum singular value is based on a similar argument, but requires somewhat more technical work, so that we defer it to the Appendix (Section 6.6). □

6.3 Covariance matrices from sub-Gaussian ensembles

Various aspects of our development thus far have crucially exploited different properties of the Gaussian distribution, especially our use of the Gaussian comparison inequalities. In this section, we show how a somewhat different approach—namely, discretization and tail bounds—can be used to establish analogous bounds for general sub-Gaussian random matrices, albeit with poorer control of the constants.

In particular, let us assume that the random vector $x_i \in \mathbb{R}^d$ is zero-mean, and sub-Gaussian with parameter at most σ, by which we mean that, for each fixed $v \in \mathbb{S}^{d-1}$,

$$\mathbb{E}[e^{\lambda\langle v,\, x_i\rangle}] \leq e^{\frac{\lambda^2\sigma^2}{2}} \qquad \text{for all } \lambda \in \mathbb{R}. \tag{6.18}$$

Equivalently stated, we assume that the scalar random variable $\langle v, x_i\rangle$ is zero-mean and sub-Gaussian with parameter at most σ. (See Chapter 2 for an in-depth discussion of sub-Gaussian variables.) Let us consider some examples to illustrate:

(a) Suppose that the matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ has i.i.d. entries, where each entry x_{ij} is zero-mean and sub-Gaussian with parameter $\sigma = 1$. Examples include the standard Gaussian ensemble ($x_{ij} \sim \mathcal{N}(0, 1)$), the Rademacher ensemble ($x_{ij} \in \{-1, +1\}$ equiprobably), and, more generally, any zero-mean distribution supported on the interval $[-1, +1]$. In all of these cases, for any vector $v \in \mathbb{S}^{d-1}$, the random variable $\langle v, x_i\rangle$ is sub-Gaussian with parameter at most σ, using the i.i.d. assumption on the entries of $x_i \in \mathbb{R}^d$, and standard properties of sub-Gaussian variables.

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press

(b) Now suppose that $x_i \sim \mathcal{N}(0, \mathbf{\Sigma})$. For any $v \in \mathbb{S}^{d-1}$, we have $\langle v,\, x_i\rangle \sim \mathcal{N}(0, v^{\mathrm{T}}\mathbf{\Sigma} v)$. Since $v^{\mathrm{T}}\mathbf{\Sigma} v \le |||\mathbf{\Sigma}|||_2$, we conclude that x_i is sub-Gaussian with parameter at most $\sigma^2 = |||\mathbf{\Sigma}|||_2$.

When the random matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ is formed by drawing each row $x_i \in \mathbb{R}^d$ in an i.i.d. manner from a σ-sub-Gaussian distribution, then we say that $\mathbf{X}$ is a sample from a *row-wise σ-sub-Gaussian ensemble*. For any such random matrix, we have the following result:

Theorem 6.5 *There are universal constants $\{c_j\}_{j=0}^{3}$ such that, for any row-wise σ-sub-Gaussian random matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, the sample covariance $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^{\mathrm{T}}$ satisfies the bounds*

$$\mathbb{E}[e^{\lambda |||\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}|||_2}] \le e^{c_0 \frac{\lambda^2\sigma^4}{n} + 4d} \qquad \text{for all } |\lambda| < \frac{n}{64e^2\sigma^2}, \tag{6.19a}$$

and hence

$$\mathbb{P}\left[\frac{|||\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}|||_2}{\sigma^2} \ge c_1\left\{\sqrt{\frac{d}{n}} + \frac{d}{n}\right\} + \delta\right] \le c_2 e^{-c_3 n \min\{\delta, \delta^2\}} \qquad \text{for all } \delta \ge 0. \tag{6.19b}$$

Remarks: Given the bound (6.19a) on the moment generating function of the random variable $|||\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}|||_2$, the tail bound (6.19b) is a straightforward consequence of the Chernoff technique (see Chapter 2). When $\mathbf{\Sigma} = \mathbf{I}_d$ and each x_i is sub-Gaussian with parameter $\sigma = 1$, the tail bound (6.19b) implies that

$$|||\widehat{\mathbf{\Sigma}} - \mathbf{I}_d|||_2 \precsim \sqrt{\frac{d}{n}} + \frac{d}{n}$$

with high probability. For $n \ge d$, this bound implies that the singular values of $\mathbf{X}/\sqrt{n}$ satisfy the sandwich relation

$$1 - c'\sqrt{\frac{d}{n}} \le \frac{\sigma_{\min}(\mathbf{X})}{\sqrt{n}} \le \frac{\sigma_{\max}(\mathbf{X})}{\sqrt{n}} \le 1 + c'\sqrt{\frac{d}{n}}, \tag{6.20}$$

for some universal constant $c' > 1$. It is worth comparing this result to the earlier bounds (6.10), applicable to the special case of a standard Gaussian matrix. The bound (6.20) has a qualitatively similar form, except that the constant c' is larger than one.

Proof For notational convenience, we introduce the shorthand $\mathbf{Q} := \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}$. Recall from Section 6.1 the variational representation $|||\mathbf{Q}|||_2 = \max_{v\in\mathbb{S}^{d-1}} |\langle v,\, \mathbf{Q}v\rangle|$. We first reduce the supremum to a finite maximum via a discretization argument (see Chapter 5). Let $\{v^1, \ldots, v^N\}$ be a $\frac{1}{8}$-covering of the sphere $\mathbb{S}^{d-1}$ in the Euclidean norm; from Example 5.8, there exists such a covering with $N \le 17^d$ vectors. Given any $v \in \mathbb{S}^{d-1}$, we can write $v = v^j + \Delta$ for some v^j in the cover, and an error vector Δ such that $\|\Delta\|_2 \le \frac{1}{8}$, and hence

$$\langle v,\, \mathbf{Q}v\rangle = \langle v^j, \mathbf{Q}v^j\rangle + 2\langle \Delta, \mathbf{Q}v^j\rangle + \langle \Delta,\, \mathbf{Q}\Delta\rangle\,.$$

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press

Applying the triangle inequality and the definition of operator norm yields

$$\begin{aligned}|\langle v,\, \mathbf{Q}v\rangle| &\le |\langle v^j, \mathbf{Q}v^j\rangle| + 2\|\Delta\|_2\, |||\mathbf{Q}|||_2\, \|v^j\|_2 + |||\mathbf{Q}|||_2\, \|\Delta\|_2^2 \\ &\le |\langle v^j, \mathbf{Q}v^j\rangle| + \tfrac{1}{4}|||\mathbf{Q}|||_2 + \tfrac{1}{64}|||\mathbf{Q}|||_2 \\ &\le |\langle v^j, \mathbf{Q}v^j\rangle| + \tfrac{1}{2}|||\mathbf{Q}|||_2.\end{aligned}$$

Rearranging and then taking the supremum over $v \in \mathbb{S}^{d-1}$, and the associated maximum over $j \in \{1, 2, \ldots, N\}$, we obtain

$$|||\mathbf{Q}|||_2 = \max_{v\in\mathbb{S}^{d-1}} |\langle v,\, \mathbf{Q}v\rangle| \le 2 \max_{j=1,\ldots,N} |\langle v^j, \mathbf{Q}v^j\rangle|.$$

Consequently, we have

$$\mathbb{E}[e^{\lambda|||\mathbf{Q}|||_2}] \le \mathbb{E}\left[\exp\left(2\lambda \max_{j=1,\ldots,N} |\langle v^j, \mathbf{Q}v^j\rangle|\right)\right] \le \sum_{j=1}^{N}\{\mathbb{E}[e^{2\lambda\langle v^j,\mathbf{Q}v^j\rangle}] + \mathbb{E}[e^{-2\lambda\langle v^j,\mathbf{Q}v^j\rangle}]\}. \tag{6.21}$$

Next we claim that for any fixed unit vector $u \in \mathbb{S}^{d-1}$,

$$\mathbb{E}[e^{t\langle u,\, \mathbf{Q}u\rangle}] \le e^{512\frac{t^2}{n}e^4\sigma^4} \qquad \text{for all } |t| \le \tfrac{n}{32e^2\sigma^2}. \tag{6.22}$$

We take this bound as given for the moment, and use it to complete the theorem's proof. For each vector v^j in the covering set, we apply the bound (6.22) twice—once with $t = 2\lambda$ and once with $t = -2\lambda$. Combining the resulting bounds with inequality (6.21), we find that

$$\mathbb{E}[e^{\lambda|||\mathbf{Q}|||_2}] \le 2Ne^{2048\frac{\lambda^2}{n}e^4\sigma^4} \le e^{c_0\frac{\lambda^2\sigma^4}{n}+4d},$$

valid for all $|\lambda| < \frac{n}{64e^2\sigma^2}$, where the final step uses the fact that $2(17^d) \le e^{4d}$. Having established the moment generating function bound (6.19a), the tail bound (6.19b) follows as a consequence of Proposition 2.9.

Proof of the bound (6.22)*:* The only remaining detail is to prove the bound (6.22). By the definition of $\mathbf{Q}$ and the i.i.d. assumption, we have

$$\mathbb{E}[e^{t\langle u,\, \mathbf{Q}u\rangle}] = \prod_{i=1}^{n} \mathbb{E}[e^{\frac{t}{n}\{\langle x_i,\, u\rangle^2 - \langle u,\, \mathbf{\Sigma}u\rangle\}}] = (\mathbb{E}[e^{\frac{t}{n}\{\langle x_1,\, u\rangle^2 - \langle u,\, \mathbf{\Sigma}u\rangle\}}])^n. \tag{6.23}$$

Letting $\varepsilon \in \{-1, +1\}$ denote a Rademacher variable, independent of x_1, a standard symmetrization argument (see Proposition 4.11) implies that

$$\begin{aligned}\mathbb{E}_{x_1}[e^{\frac{t}{n}\{\langle x_1,\, u\rangle^2 - \langle u,\, \mathbf{\Sigma}u\rangle\}}] \le \mathbb{E}_{x_1,\varepsilon}[e^{\frac{2t}{n}\varepsilon\langle x_1,\, u\rangle^2}] &\overset{\text{(i)}}{=} \sum_{k=0}^{\infty} \frac{1}{k!}\left(\frac{2t}{n}\right)^k \mathbb{E}[\varepsilon^k \langle x_1,\, u\rangle^{2k}] \\ &\overset{\text{(ii)}}{=} 1 + \sum_{\ell=1}^{\infty} \frac{1}{(2\ell)!}\left(\frac{2t}{n}\right)^{2\ell} \mathbb{E}[\langle x_1,\, u\rangle^{4\ell}],\end{aligned}$$

where step (i) follows by the power-series expansion of the exponential, and step (ii) follows since ε and x_1 are independent, and all odd moments of the Rademacher term vanish. By property (III) in Theorem 2.6 on equivalent characterizations of sub-Gaussian variables, we

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press

are guaranteed that

$$\mathbb{E}[\langle x_1, u\rangle^{4\ell}] \leq \frac{(4\ell)!}{2^{2\ell}(2\ell)!}(\sqrt{8}e\sigma)^{4\ell} \qquad \text{for all } \ell = 1, 2, \ldots,$$

and hence

$$\begin{aligned}\mathbb{E}_{x_1}[e^{\frac{t}{n}\{\langle x_1, u\rangle^2 - \langle u, \Sigma u\rangle\}}] &\leq 1 + \sum_{\ell=1}^{\infty} \frac{1}{(2\ell)!}\left(\frac{2t}{n}\right)^{2\ell} \frac{(4\ell)!}{2^{2\ell}(2\ell)!}(\sqrt{8}e\sigma)^{4\ell} \\ &\leq 1 + \sum_{\ell=1}^{\infty} \Big(\underbrace{\frac{16t}{n}e^2\sigma^2}_{f(t)}\Big)^{2\ell},\end{aligned}$$

where we have used the fact that $(4\ell)! \leq 2^{2\ell}[(2\ell)!]^2$. As long as $f(t) := \frac{16t}{n}e^2\sigma^2 < \frac{1}{2}$, we can write

$$1 + \sum_{\ell=1}^{\infty}[f^2(t)]^{\ell} \overset{\text{(i)}}{=} \frac{1}{1 - f^2(t)} \overset{\text{(ii)}}{\leq} \exp(2f^2(t)),$$

where step (i) follows by summing the geometric series, and step (ii) follows because $\frac{1}{1-a} \leq e^{2a}$ for all $a \in [0, \frac{1}{2}]$. Putting together the pieces and combining with our earlier bound (6.23), we have shown that $\mathbb{E}[e^{t\langle u, \mathbf{Q}u\rangle}] \leq e^{2nf^2(t)}$, valid for all $|t| < \frac{n}{32e^2\sigma^2}$, which establishes the claim (6.22). □

6.4 Bounds for general matrices

The preceding sections were devoted to bounds applicable to sample covariances under Gaussian or sub-Gaussian tail conditions. This section is devoted to developing extensions to more general tail conditions. In order to do so, it is convenient to introduce some more general methodology that applies not only to sample covariance matrices, but also to more general random matrices. The main results in this section are Theorems 6.15 and 6.17, which are (essentially) matrix-based analogs of our earlier Hoeffding and Bernstein bounds for random variables. Before proving these results, we develop some useful matrix-theoretic generalizations of ideas from Chapter 2, including various types of tail conditions, as well as decompositions for the moment generating function for independent random matrices.

6.4.1 Background on matrix analysis

We begin by introducing some additional background on matrix-valued functions. Recall the class $\mathcal{S}^{d\times d}$ of symmetric $d \times d$ matrices. Any function $f\colon \mathbb{R} \to \mathbb{R}$ can be extended to a map from the set $\mathcal{S}^{d\times d}$ to itself in the following way. Given a matrix $\mathbf{Q} \in \mathcal{S}^{d\times d}$, consider its eigendecomposition $\mathbf{Q} = \mathbf{U}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{U}$. Here the matrix $\mathbf{U} \in \mathbb{R}^{d\times d}$ is a unitary matrix, satisfying the relation $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}_d$, whereas $\mathbf{\Gamma} := \mathrm{diag}(\gamma(\mathbf{Q}))$ is a diagonal matrix specified by the vector of eigenvalues $\gamma(\mathbf{Q}) \in \mathbb{R}^d$. Using this notation, we consider the mapping from $\mathcal{S}^{d\times d}$ to itself defined via

$$\mathbf{Q} \mapsto f(\mathbf{Q}) := \mathbf{U}^{\mathrm{T}}\,\mathrm{diag}(f(\gamma_1(\mathbf{Q})), \ldots, f(\gamma_d(\mathbf{Q})))\mathbf{U}.$$

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press

In words, we apply the original function f elementwise to the vector of eigenvalues $\gamma(\mathbf{Q})$, and then rotate the resulting matrix $\operatorname{diag}(f(\gamma(\mathbf{Q})))$ back to the original coordinate system defined by the eigenvectors of $\mathbf{Q}$. By construction, this extension of f to $\mathcal{S}^{d\times d}$ is unitarily invariant, meaning that

$$f(\mathbf{V}^{\mathrm{T}}\mathbf{Q}\mathbf{V}) = \mathbf{V}^{\mathrm{T}} f(\mathbf{Q})\mathbf{V} \qquad \text{for all unitary matrices } \mathbf{V} \in \mathbb{R}^{d\times d},$$

since it affects only the eigenvalues (but not the eigenvectors) of $\mathbf{Q}$. Moreover, the eigenvalues of $f(\mathbf{Q})$ transform in a simple way, since we have

$$\gamma(f(\mathbf{Q})) = \{f(\gamma_j(\mathbf{Q})),\ j = 1, \ldots, d\}. \tag{6.24}$$

In words, the eigenvalues of the matrix $f(\mathbf{Q})$ are simply the eigenvalues of $\mathbf{Q}$ transformed by f, a result often referred to as the *spectral mapping property*.

Two functions that play a central role in our development of matrix tail bounds are the matrix exponential and the matrix logarithm. As a particular case of our construction, the matrix exponential has the power-series expansion $e^{\mathbf{Q}} = \sum_{k=0}^{\infty} \frac{\mathbf{Q}^k}{k!}$. By the spectral mapping property, the eigenvalues of $e^{\mathbf{Q}}$ are positive, so that it is a positive definite matrix for any choice of $\mathbf{Q}$. Parts of our analysis also involve the matrix logarithm; when restricted to the cone of strictly positive definite matrices, as suffices for our purposes, the matrix logarithm corresponds to the inverse of the matrix exponential.

A function f on $\mathcal{S}^{d\times d}$ is said to be *matrix monotone* if $f(\mathbf{Q}) \preceq f(\mathbf{R})$ whenever $\mathbf{Q} \preceq \mathbf{R}$. A useful property of the logarithm is that it is a matrix monotone function, a result known as the *Löwner–Heinz theorem.* By contrast, the exponential is *not* a matrix monotone function, showing that matrix monotonicity is more complex than the usual notion of monotonicity. See Exercise 6.5 for further exploration of these properties.

Finally, a useful fact is the following: if $f\colon \mathbb{R} \to \mathbb{R}$ is any continuous and non-decreasing function in the usual sense, then for any pair of symmetric matrices such that $\mathbf{Q} \preceq \mathbf{R}$, we are guaranteed that

$$\operatorname{tr}(f(\mathbf{Q})) \leq \operatorname{tr}(f(\mathbf{R})). \tag{6.25}$$

See the bibliographic section for further discussion of such *trace inequalities*.

6.4.2 Tail conditions for matrices

Given a symmetric random matrix $\mathbf{Q} \in \mathcal{S}^{d\times d}$, its polynomial moments, assuming that they exist, are the matrices defined by $\mathbb{E}[\mathbf{Q}^j]$. As shown in Exercise 6.6, the variance of $\mathbf{Q}$ is a positive semidefinite matrix given by $\operatorname{var}(\mathbf{Q}) := \mathbb{E}[\mathbf{Q}^2] - (\mathbb{E}[\mathbf{Q}])^2$. The moment generating function of a random matrix $\mathbf{Q}$ is the matrix-valued mapping $\Psi_{\mathbf{Q}}\colon \mathbb{R} \to \mathcal{S}^{d\times d}$ given by

$$\Psi_{\mathbf{Q}}(\lambda) := \mathbb{E}[e^{\lambda\mathbf{Q}}] = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}\mathbb{E}[\mathbf{Q}^k]. \tag{6.26}$$

Under suitable conditions on $\mathbf{Q}$—or equivalently, suitable conditions on the polynomial moments of $\mathbf{Q}$—it is guaranteed to be finite for all λ in an interval centered at zero. In parallel with our discussion in Chapter 2, various tail conditions are based on imposing bounds on this moment generating function. We begin with the simplest case:

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press

Definition 6.6 A zero-mean symmetric random matrix $\mathbf{Q} \in \mathcal{S}^{d\times d}$ is sub-Gaussian with matrix parameter $\mathbf{V} \in \mathcal{S}_+^{d\times d}$ if

$$\Psi_{\mathbf{Q}}(\lambda) \preceq e^{\frac{\lambda^2 \mathbf{V}}{2}} \qquad \text{for all } \lambda \in \mathbb{R}. \tag{6.27}$$

This definition is best understood by working through some simple examples.

Example 6.7 Suppose that $\mathbf{Q} = \varepsilon \mathbf{B}$ where $\varepsilon \in \{-1, +1\}$ is a Rademacher variable, and $\mathbf{B} \in \mathcal{S}^{d\times d}$ is a fixed matrix. Random matrices of this form frequently arise as the result of symmetrization arguments, as discussed at more length in the sequel. Note that we have $\mathbb{E}[\mathbf{Q}^{2k+1}] = 0$ and $\mathbb{E}[\mathbf{Q}^{2k}] = \mathbf{B}^{2k}$ for all $k = 1, 2, \ldots$, and hence

$$\mathbb{E}[e^{\lambda \mathbf{Q}}] = \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{(2k)!}\mathbf{B}^{2k} \preceq \sum_{k=1}^{\infty} \frac{1}{k!}\left(\frac{\lambda^2 \mathbf{B}^2}{2}\right)^k = e^{\frac{\lambda^2 \mathbf{B}^2}{2}},$$

showing that the sub-Gaussian condition (6.27) holds with $\mathbf{V} = \mathbf{B}^2 = \text{var}(\mathbf{Q})$. ♣

As we show in Exercise 6.7, more generally, a random matrix of the form $\mathbf{Q} = g\mathbf{B}$, where $g \in \mathbb{R}$ is a σ-sub-Gaussian variable with distribution symmetric around zero, satisfies the condition (6.27) with matrix parameter $\mathbf{V} = \sigma^2 \mathbf{B}^2$.

Example 6.8 As an extension of the previous example, consider a random matrix of the form $\mathbf{Q} = \varepsilon \mathbf{C}$, where ε is a Rademacher variable as before, and $\mathbf{C}$ is now a random matrix, independent of ε with its spectral norm bounded as $|||\mathbf{C}|||_2 \leq b$. First fixing $\mathbf{C}$ and taking expectations over the Rademacher variable, the previous example yields $\mathbb{E}_\varepsilon[e^{\lambda\varepsilon\mathbf{C}}] \preceq e^{\frac{\lambda^2}{2}\mathbf{C}^2}$. Since $|||\mathbf{C}|||_2 \leq b$, we have $e^{\frac{\lambda^2}{2}\mathbf{C}^2} \preceq e^{\frac{\lambda^2}{2}b^2 \mathbf{I}_d}$, and hence

$$\Psi_{\mathbf{Q}}(\lambda) \preceq e^{\frac{\lambda^2}{2}b^2 \mathbf{I}_d} \qquad \text{for all } \lambda \in \mathbb{R},$$

showing that $\mathbf{Q}$ is sub-Gaussian with matrix parameter $\mathbf{V} = b^2 \mathbf{I}_d$. ♣

In parallel with our treatment of scalar random variables in Chapter 2, it is natural to consider various weakenings of the sub-Gaussian requirement.

Definition 6.9 (Sub-exponential random matrices) A zero-mean random matrix is sub-exponential with parameters $(\mathbf{V}, \alpha)$ if

$$\Psi_{\mathbf{Q}}(\lambda) \preceq e^{\frac{\lambda^2 \mathbf{V}}{2}} \qquad \text{for all } |\lambda| < \tfrac{1}{\alpha}. \tag{6.28}$$

Thus, any sub-Gaussian random matrix is also sub-exponential with parameters $(\mathbf{V}, 0)$. However, there also exist sub-exponential random matrices that are not sub-Gaussian. One example is the zero-mean random matrix $\mathbf{M} = \varepsilon g^2 \mathbf{B}$, where $\varepsilon \in \{-1, +1\}$ is a Rademacher

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press

variable, the variable $g \sim \mathcal{N}(0,1)$ is independent of ε, and $\mathbf{B}$ is a fixed symmetric matrix.

The Bernstein condition for random matrices provides one useful way of certifying the sub-exponential condition:

Definition 6.10 (Bernstein's condition for matrices) A zero-mean symmetric random matrix $\mathbf{Q}$ satisfies a Bernstein condition with parameter $b > 0$ if

$$\mathbb{E}[\mathbf{Q}^j] \preceq \tfrac{1}{2} j!\, b^{j-2}\,\mathrm{var}(\mathbf{Q}) \qquad \text{for } j = 3, 4, \ldots. \tag{6.29}$$

We note that (a stronger form of) Bernstein's condition holds whenever the matrix $\mathbf{Q}$ has a bounded operator norm—say $|||\mathbf{Q}|||_2 \leq b$ almost surely. In this case, it can be shown (see Exercise 6.9) that

$$\mathbb{E}[\mathbf{Q}^j] \preceq b^{j-2}\,\mathrm{var}(\mathbf{Q}) \qquad \text{for all } j = 3, 4, \ldots. \tag{6.30}$$

Exercise 6.11 gives an example of a random matrix with unbounded operator norm for which Bernstein's condition holds.

The following lemma shows how the general Bernstein condition (6.29) implies the sub-exponential condition. More generally, the argument given here provides an explicit bound on the moment generating function:

Lemma 6.11 *For any symmetric zero-mean random matrix satisfying the Bernstein condition* (6.29), *we have*

$$\Psi_{\mathbf{Q}}(\lambda) \preceq \exp\left(\frac{\lambda^2\,\mathrm{var}(\mathbf{Q})}{2(1 - b|\lambda|)}\right) \qquad \textit{for all } |\lambda| < \frac{1}{b}. \tag{6.31}$$

Proof Since $\mathbb{E}[\mathbf{Q}] = 0$, applying the definition of the matrix exponential for a suitably small $\lambda \in \mathbb{R}$ yields

$$\begin{aligned}
\mathbb{E}[e^{\lambda\mathbf{Q}}] &= \mathbf{I}_d + \frac{\lambda^2\,\mathrm{var}(\mathbf{Q})}{2} + \sum_{j=3}^{\infty} \frac{\lambda^j \mathbb{E}[\mathbf{Q}^j]}{j!} \\
&\overset{\text{(i)}}{\preceq} \mathbf{I}_d + \frac{\lambda^2\,\mathrm{var}(\mathbf{Q})}{2}\left\{\sum_{j=0}^{\infty} |\lambda|^j b^j\right\} \\
&\overset{\text{(ii)}}{=} \mathbf{I}_d + \frac{\lambda^2\,\mathrm{var}(\mathbf{Q})}{2(1 - b|\lambda|)} \\
&\overset{\text{(iii)}}{\preceq} \exp\left(\frac{\lambda^2\,\mathrm{var}(\mathbf{Q})}{2(1 - b|\lambda|)}\right),
\end{aligned}$$

where step (i) applies the Bernstein condition, step (ii) is valid for any $|\lambda| < 1/b$, a choice for which the geometric series is summable, and step (iii) follows from the matrix inequality

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press

$\mathbf{I}_d + \mathbf{A} \preceq e^{\mathbf{A}}$, which is valid for any symmetric matrix $\mathbf{A}$. (See Exercise 6.4 for more discussion of this last property.) □

6.4.3 Matrix Chernoff approach and independent decompositions

The Chernoff approach to tail bounds, as discussed in Chapter 2, is based on controlling the moment generating function of a random variable. In this section, we begin by showing that the trace of the matrix moment generating function (6.26) plays a similar role in bounding the operator norm of random matrices.

Lemma 6.12 (Matrix Chernoff technique) *Let* $\mathbf{Q}$ *be a zero-mean symmetric random matrix whose moment generating function* $\Psi_{\mathbf{Q}}$ *exists in an open interval* $(-a, a)$*. Then for any* $\delta > 0$*, we have*

$$\mathbb{P}[\gamma_{\max}(\mathbf{Q}) \geq \delta] \leq \operatorname{tr}(\Psi_{\mathbf{Q}}(\lambda))e^{-\lambda\delta} \qquad \textit{for all } \lambda \in [0, a), \tag{6.32}$$

where $\operatorname{tr}(\cdot)$ *denotes the trace operator on matrices. Similarly, we have*

$$\mathbb{P}[|||\mathbf{Q}|||_2 \geq \delta] \leq 2\operatorname{tr}(\Psi_{\mathbf{Q}}(\lambda))e^{-\lambda\delta} \qquad \textit{for all } \lambda \in [0, a). \tag{6.33}$$

Proof For each $\lambda \in [0, a)$, we have

$$\mathbb{P}[\gamma_{\max}(\mathbf{Q}) \geq \delta] = \mathbb{P}[e^{\gamma_{\max}(\lambda\mathbf{Q})} \geq e^{\lambda\delta}] \overset{\text{(i)}}{=} \mathbb{P}[\gamma_{\max}(e^{\lambda\mathbf{Q}}) \geq e^{\lambda\delta}], \tag{6.34}$$

where step (i) uses the functional calculus relating the eigenvalues of $\lambda\mathbf{Q}$ to those of $e^{\lambda\mathbf{Q}}$. Applying Markov's inequality yields

$$\mathbb{P}[\gamma_{\max}(e^{\lambda\mathbf{Q}}) \geq e^{\lambda\delta}] \leq \mathbb{E}[\gamma_{\max}(e^{\lambda\mathbf{Q}})]e^{-\lambda\delta} \overset{\text{(i)}}{\leq} \mathbb{E}[\operatorname{tr}(e^{\lambda\mathbf{Q}})]e^{-\lambda\delta}. \tag{6.35}$$

Here inequality (i) uses the upper bound $\gamma_{\max}(e^{\lambda\mathbf{Q}}) \leq \operatorname{tr}(e^{\lambda\mathbf{Q}})$, which holds since $e^{\lambda\mathbf{Q}}$ is positive definite. Finally, since trace and expectation commute, we have

$$\mathbb{E}[\operatorname{tr}(e^{\lambda\mathbf{Q}})] = \operatorname{tr}(\mathbb{E}[e^{\lambda\mathbf{Q}}]) = \operatorname{tr}(\Psi_{\mathbf{Q}}(\lambda)).$$

Note that the same argument can be applied to bound the event $\gamma_{\max}(-\mathbf{Q}) \geq \delta$, or equivalently the event $\gamma_{\min}(\mathbf{Q}) \leq -\delta$. Since $|||\mathbf{Q}|||_2 = \max\{\gamma_{\max}(\mathbf{Q}), |\gamma_{\min}(\mathbf{Q})|\}$, the tail bound on the operator norm (6.33) follows. □

An important property of independent random variables is that the moment generating function of their sum can be decomposed as the product of the individual moment generating functions. For random matrices, this type of decomposition is no longer guaranteed to hold with equality, essentially because matrix products need not commute. However, for independent random matrices, it is nonetheless possible to establish an upper bound in terms of the trace of the product of moment generating functions, as we now show.

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press

Lemma 6.13 *Let $\mathbf{Q}_1, \ldots, \mathbf{Q}_n$ be independent symmetric random matrices whose moment generating functions exist for all $\lambda \in I$, and define the sum $\mathbf{S}_n := \sum_{i=1}^n \mathbf{Q}_i$. Then*

$$\operatorname{tr}(\Psi_{\mathbf{S}_n}(\lambda)) \leq \operatorname{tr}\Big(e^{\sum_{i=1}^n \log \Psi_{\mathbf{Q}_i}(\lambda)}\Big) \qquad \textit{for all } \lambda \in I. \tag{6.36}$$

Remark: In conjunction with Lemma 6.12, this lemma provides an avenue for obtaining tail bounds on the operator norm of sums of independent random matrices. In particular, if we apply the upper bound (6.33) to the random matrix $\mathbf{S}_n/n$, we find that

$$\mathbb{P}\left[\left|\!\left|\!\left|\frac{1}{n}\sum_{i=1}^n \mathbf{Q}_i\right|\!\right|\!\right|_2 \geq \delta\right] \leq 2\operatorname{tr}\Big(e^{\sum_{i=1}^n \log \Psi_{\mathbf{Q}_i}(\lambda)}\Big)\, e^{-\lambda n \delta} \qquad \text{for all } \lambda \in [0, a). \tag{6.37}$$

Proof In order to prove this lemma, we require the following result due to Lieb (1973): for any fixed matrix $\mathbf{H} \in \mathcal{S}^{d\times d}$, the function $f\colon \mathcal{S}_+^{d\times d} \to \mathbb{R}$ given by

$$f(\mathbf{A}) := \operatorname{tr}(e^{\mathbf{H}+\log(\mathbf{A})})$$

is concave. Introducing the shorthand notation $G(\lambda) := \operatorname{tr}(\Psi_{\mathbf{S}_n}(\lambda))$, we note that, by linearity of trace and expectation, we have

$$G(\lambda) = \operatorname{tr}\Big(\mathbb{E}[e^{\lambda \mathbf{S}_{n-1}+\log \exp(\lambda \mathbf{Q}_n)}]\Big) = \mathbb{E}_{\mathbf{S}_{n-1}}\, \mathbb{E}_{\mathbf{Q}_n}[\operatorname{tr}(e^{\lambda \mathbf{S}_{n-1}+\log \exp(\lambda \mathbf{Q}_n)})].$$

Using concavity of the function f with $\mathbf{H} = \lambda \mathbf{S}_{n-1}$ and $\mathbf{A} = e^{\lambda \mathbf{Q}_n}$, Jensen's inequality implies that

$$\mathbb{E}_{\mathbf{Q}_n}[\operatorname{tr}(e^{\lambda \mathbf{S}_{n-1}+\log \exp(\lambda \mathbf{Q}_n)})] \leq \operatorname{tr}(e^{\lambda \mathbf{S}_{n-1}+\log \mathbb{E}_{\mathbf{Q}_n} \exp(\lambda \mathbf{Q}_n)}),$$

so that we have shown that $G(\lambda) \leq \mathbb{E}_{\mathbf{S}_{n-1}}[\operatorname{tr}(e^{\lambda \mathbf{S}_{n-1}+\log \Psi_{\mathbf{Q}_n}(\lambda)})]$.

We now recurse this argument, in particular peeling off the term involving $\mathbf{Q}_{n-1}$, so that we have

$$G(\lambda) \leq \mathbb{E}_{\mathbf{S}_{n-2}} \mathbb{E}_{\mathbf{Q}_{n-1}}\big[\operatorname{tr}(e^{\lambda \mathbf{S}_{n-2}+\log \Psi_{\mathbf{Q}_n}(\lambda)+\log \exp(\lambda \mathbf{Q}_{n-1})})\big].$$

We again exploit the concavity of the function f, this time with the choices $\mathbf{H} = \lambda \mathbf{S}_{n-2} + \log \Psi_{\mathbf{Q}_n}(\lambda)$ and $\mathbf{A} = e^{\lambda \mathbf{Q}_{n-1}}$, thereby finding that

$$G(\lambda) \leq \mathbb{E}_{\mathbf{S}_{n-2}}\big[\operatorname{tr}(e^{\lambda \mathbf{S}_{n-2}+\log \Psi_{\mathbf{Q}_{n-1}}(\lambda)+\log \Psi_{\mathbf{Q}_n}(\lambda)})\big].$$

Continuing in this manner completes the proof of the claim. □

In many cases, our goal is to bound the maximum eigenvalue (or operator norm) of sums of centered random matrices of the form $\mathbf{Q}_i = \mathbf{A}_i - \mathbb{E}[\mathbf{A}_i]$. In this and other settings, it is often convenient to perform an additional symmetrization step, so that we can deal instead with matrices $\widetilde{\mathbf{Q}}_i$ that are guaranteed to have distribution symmetric around zero (meaning that $\widetilde{\mathbf{Q}}_i$ and $-\widetilde{\mathbf{Q}}_i$ follow the same distribution).

Example 6.14 (Rademacher symmetrization for random matrices) Let $\{\mathbf{A}_i\}_{i=1}^n$ be a sequence of independent symmetric random matrices, and suppose that our goal is to bound the maximum eigenvalue of the matrix sum $\sum_{i=1}^n (\mathbf{A}_i - \mathbb{E}[\mathbf{A}_i])$. Since the maximum eigenvalue can be represented as the supremum of an empirical process, the symmetrization techniques from Chapter 4 can be used to reduce the problem to one involving the new matrices

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press

$\widetilde{\mathbf{Q}}_i = \varepsilon_i \mathbf{A}_i$, where ε_i is an independent Rademacher variable. Let us now work through this reduction. By Markov's inequality, we have

$$\mathbb{P}\left[\gamma_{\max}\left(\sum_{i=1}^n \{\mathbf{A}_i - \mathbb{E}[\mathbf{A}_i]\}\right) \geq \delta\right] \leq \mathbb{E}[e^{\lambda\gamma_{\max}(\sum_{i=1}^n \{\mathbf{A}_i - \mathbb{E}[\mathbf{A}_i]\})}]e^{-\lambda\delta}.$$

By the variational representation of the maximum eigenvalue, we have

$$\begin{aligned}\mathbb{E}[e^{\lambda\gamma_{\max}(\sum_{i=1}^n \{\mathbf{A}_i - \mathbb{E}[\mathbf{A}_i]\})}] &= \mathbb{E}\left[\exp\left(\lambda \sup_{\|u\|_2=1}\left\langle u, \left(\sum_{i=1}^n (\mathbf{A}_i - \mathbb{E}[\mathbf{A}_i])\right)u\right\rangle\right)\right] \\ &\overset{\text{(i)}}{\leq} \mathbb{E}\left[\exp\left(2\lambda \sup_{\|u\|_2=1}\left\langle u, \left(\sum_{i=1}^n \varepsilon_i \mathbf{A}_i\right)u\right\rangle\right)\right] \\ &= \mathbb{E}[e^{2\lambda\gamma_{\max}(\sum_{i=1}^n \varepsilon_i \mathbf{A}_i)}] \\ &\overset{\text{(ii)}}{=} \mathbb{E}[\gamma_{\max}(e^{2\lambda\sum_{i=1}^n \varepsilon_i \mathbf{A}_i})],\end{aligned}$$

where inequality (i) makes use of the symmetrization inequality from Proposition 4.11(b) with $\Phi(t) = e^{\lambda t}$, and step (ii) uses the spectral mapping property (6.24). Continuing on, we have

$$\mathbb{E}[\gamma_{\max}(e^{2\lambda\sum_{i=1}^n \varepsilon_i \mathbf{A}_i})] \leq \operatorname{tr}\left(\mathbb{E}[e^{2\lambda\sum_{i=1}^n \varepsilon_i \mathbf{A}_i}]\right) \leq \operatorname{tr}\left(e^{\sum_{i=1}^n \log \Psi_{\widetilde{\mathbf{Q}}_i}(2\lambda)}\right),$$

where the final step follows from applying Lemma 6.13 to the symmetrized matrices $\widetilde{\mathbf{Q}}_i = \varepsilon_i \mathbf{A}_i$. Consequently, apart from the factor of 2, we may assume without loss of generality when bounding maximum eigenvalues that our matrices have a distribution symmetric around zero. ♣

6.4.4 *Upper tail bounds for random matrices*

We now have collected the ingredients necessary for stating and proving various tail bounds for the deviations of sums of zero-mean independent random matrices.

Sub-Gaussian case

We begin with a tail bound for sub-Gaussian random matrices. It provides an approximate analog of the Hoeffding-type tail bound for random variables (Proposition 2.5).

Theorem 6.15 (Hoeffding bound for random matrices) *Let $\{\mathbf{Q}_i\}_{i=1}^n$ be a sequence of zero-mean independent symmetric random matrices that satisfy the sub-Gaussian condition with parameters $\{\mathbf{V}_i\}_{i=1}^n$. Then for all $\delta > 0$, we have the upper tail bound*

$$\mathbb{P}\left[\left|\!\left|\!\left|\frac{1}{n}\sum_{i=1}^n \mathbf{Q}_i\right|\!\right|\!\right|_2 \geq \delta\right] \leq 2\operatorname{rank}\left(\sum_{i=1}^n \mathbf{V}_i\right)e^{-\frac{n\delta^2}{2\sigma^2}} \leq 2de^{-\frac{n\delta^2}{2\sigma^2}}, \tag{6.38}$$

where $\sigma^2 = |\!|\!|\frac{1}{n}\sum_{i=1}^n \mathbf{V}_i|\!|\!|_2$.

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press

Proof We first prove the claim in the case when $\mathbf{V} := \sum_{i=1}^n \mathbf{V}_i$ is full-rank, and then show how to prove the general case. From Lemma 6.13, it suffices to upper bound $\operatorname{tr}\Big(e^{\sum_{i=1}^n \log \Psi_{\mathbf{Q}_i}(\lambda)}\Big)$. From Definition 6.6, the assumed sub-Gaussianity, and the monotonicity of the matrix logarithm, we have

$$\sum_{i=1}^n \log \Psi_{\mathbf{Q}_i}(\lambda) \preceq \frac{\lambda^2}{2} \sum_{i=1}^n \mathbf{V}_i,$$

where we have used the fact that the logarithm is matrix monotone. Now since the exponential is an increasing function, the trace bound (6.25) implies that

$$\operatorname{tr}\Big(e^{\sum_{i=1}^n \log \Psi_{\mathbf{Q}_i}(\lambda)}\Big) \le \operatorname{tr}\Big(e^{\frac{\lambda^2}{2}\sum_{i=1}^n \mathbf{V}_i}\Big).$$

This upper bound, when combined with the matrix Chernoff bound (6.37), yields

$$\mathbb{P}\Bigg[\Bigg|\!\Bigg|\!\Bigg|\frac{1}{n}\sum_{i=1}^n \mathbf{Q}_i\Bigg|\!\Bigg|\!\Bigg|_2 \ge \delta\Bigg] \le 2\operatorname{tr}\Big(e^{\frac{\lambda^2}{2}\sum_{i=1}^n \mathbf{V}_i}\Big)e^{-\lambda n\delta}.$$

For any d-dimensional symmetric matrix $\mathbf{R}$, we have $\operatorname{tr}(e^{\mathbf{R}}) \le d e^{|||\mathbf{R}|||_2}$. Applying this inequality to the matrix $\mathbf{R} = \frac{\lambda^2}{2}\sum_{i=1}^n \mathbf{V}_i$, for which we have $|||\mathbf{R}|||_2 = \frac{\lambda^2}{2} n\sigma^2$, yields the bound

$$\mathbb{P}\Bigg[\Bigg|\!\Bigg|\!\Bigg|\frac{1}{n}\sum_{i=1}^n \mathbf{Q}_i\Bigg|\!\Bigg|\!\Bigg|_2 \ge \delta\Bigg] \le 2d e^{\frac{\lambda^2}{2} n\sigma^2 - \lambda n\delta}.$$

This upper bound holds for all $\lambda \ge 0$ and setting $\lambda = \delta/\sigma^2$ yields the claim.

Now suppose that the matrix $\mathbf{V} := \sum_{i=1}^n \mathbf{V}_i$ is not full-rank, say of rank $r < d$. In this case, an eigendecomposition yields $\mathbf{V} = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}}$, where $\mathbf{U} \in \mathbb{R}^{d\times r}$ has orthonormal columns. Introducing the shorthand $\mathbf{Q} := \sum_{i=1}^n \mathbf{Q}_i$, the r-dimensional matrix $\widetilde{\mathbf{Q}} = \mathbf{U}^{\mathrm{T}}\mathbf{Q}\mathbf{U}$ then captures all randomness in $\mathbf{Q}$, and in particular we have $|||\widetilde{\mathbf{Q}}|||_2 = |||\mathbf{Q}|||_2$. We can thus apply the same argument to bound $|||\widetilde{\mathbf{Q}}|||_2$, leading to a pre-factor of r instead of d. □

An important fact is that inequality (6.38) also implies an analogous bound for general independent but potentially non-symmetric and/or non-square matrices, with d replaced by $(d_1 + d_2)$. More specifically, a problem involving general zero-mean random matrices $\mathbf{A}_i \in \mathbb{R}^{d_1\times d_2}$ can be transformed to a symmetric version by defining the $(d_1 + d_2)$-dimensional square matrices

$$\mathbf{Q}_i := \begin{bmatrix} \mathbf{0}_{d_1\times d_1} & \mathbf{A}_i \\ \mathbf{A}_i^{\mathrm{T}} & \mathbf{0}_{d_2\times d_2} \end{bmatrix}, \tag{6.39}$$

and imposing some form of moment generating function bound—for instance, the sub-Gaussian condition (6.27)—on the symmetric matrices $\mathbf{Q}_i$. See Exercise 6.10 for further details.

A significant feature of the tail bound (6.38) is the appearance of either the rank or the dimension d in front of the exponent. In certain cases, this dimension-dependent factor is superfluous, and leads to sub-optimal bounds. However, it cannot be avoided in general. The following example illustrates these two extremes.

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press

Example 6.16 (Looseness/sharpness of Theorem 6.15) For simplicity, let us consider examples with $n = d$. For each $i = 1, 2, \ldots, d$, let $\mathbf{E}_i \in \mathcal{S}^{d\times d}$ denote the diagonal matrix with 1 in position (i, i), and 0s elsewhere. Define $\mathbf{Q}_i = y_i \mathbf{E}_i$, where $\{y_i\}_{i=1}^n$ is an i.i.d. sequence of 1-sub-Gaussian variables. Two specific cases to keep in mind are Rademacher variables $\{\varepsilon_i\}_{i=1}^n$, and $\mathcal{N}(0, 1)$ variables $\{g_i\}_{i=1}^n$.

For any such choice of sub-Gaussian variables, a calculation similar to that of Example 6.7 shows that each $\mathbf{Q}_i$ satisfies the sub-Gaussian bound (6.27) with $\mathbf{V}_i = \mathbf{E}_i$, and hence $\sigma^2 = |||\frac{1}{d}\sum_{i=1}^d \mathbf{V}_i|||_2 = 1/d$. Consequently, an application of Theorem 6.15 yields the tail bound

$$\mathbb{P}\left[\left|\left|\left|\frac{1}{d}\sum_{i=1}^d \mathbf{Q}_i\right|\right|\right|_2 \geq \delta\right] \leq 2de^{-\frac{d^2\delta^2}{2}} \qquad \text{for all } \delta > 0, \tag{6.40}$$

which implies that $|||\frac{1}{d}\sum_{j=1}^d \mathbf{Q}_j|||_2 \precsim \frac{\sqrt{2\log(2d)}}{d}$ with high probability. On the other hand, an explicit calculation shows that

$$\left|\left|\left|\frac{1}{d}\sum_{i=1}^n \mathbf{Q}_i\right|\right|\right|_2 = \max_{i=1,\ldots,d} \frac{|y_i|}{d}. \tag{6.41}$$

Comparing the exact result (6.41) with the bound (6.40) yields a range of behavior. At one extreme, for i.i.d. Rademacher variables $y_i = \varepsilon_i \in \{-1, +1\}$, we have $|||\frac{1}{d}\sum_{i=1}^n \mathbf{Q}_i|||_2 = 1/d$, showing that the bound (6.40) is off by the order $\sqrt{\log d}$. On the other hand, for i.i.d. Gaussian variables $y_i = g_i \sim \mathcal{N}(0, 1)$, we have

$$\left|\left|\left|\frac{1}{d}\sum_{i=1}^d \mathbf{Q}_i\right|\right|\right|_2 = \max_{i=1,\ldots,d} \frac{|g_i|}{d} \simeq \frac{\sqrt{2\log d}}{d},$$

using the fact that the maximum of d i.i.d. $\mathcal{N}(0, 1)$ variables scales as $\sqrt{2\log d}$. Consequently, Theorem 6.15 cannot be improved for this class of random matrices. ♣

Bernstein-type bounds for random matrices

We now turn to bounds on random matrices that satisfy sub-exponential tail conditions, in particular of the Bernstein form (6.29).

Theorem 6.17 (Bernstein bound for random matrices) *Let $\{\mathbf{Q}_i\}_{i=1}^n\}$ be a sequence of independent, zero-mean, symmetric random matrices that satisfy the Bernstein condition* (6.29) *with parameter $b > 0$. Then for all $\delta \geq 0$, the operator norm satisfies the tail bound*

$$\mathbb{P}\left[\frac{1}{n}\left|\left|\left|\sum_{i=1}^n \mathbf{Q}_i\right|\right|\right|_2 \geq \delta\right] \leq 2\,\mathrm{rank}\left(\sum_{i=1}^n \mathrm{var}(\mathbf{Q}_i)\right)\exp\left\{-\frac{n\delta^2}{2(\sigma^2 + b\delta)}\right\}, \tag{6.42}$$

where $\sigma^2 := \frac{1}{n}|||\sum_{j=1}^n \mathrm{var}(\mathbf{Q}_j)|||_2$.

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press

Proof By Lemma 6.13, we have $\operatorname{tr}(\Psi_{\mathbf{S}_n}(\lambda)) \leq \operatorname{tr}\left(e^{\sum_{i=1}^n \log \Psi_{\mathbf{Q}_i}(\lambda)}\right)$. By Lemma 6.11, the Bernstein condition combined with matrix monotonicity of the logarithm yields the bound $\log \Psi_{\mathbf{Q}_i}(\lambda) \preceq \frac{\lambda^2 \operatorname{var}(\mathbf{Q}_i)}{1-b|\lambda|}$ for any $|\lambda| < \frac{1}{b}$. Putting together the pieces yields

$$\operatorname{tr}\left(e^{\sum_{i=1}^n \log \Psi_{\mathbf{Q}_i}(\lambda)}\right) \leq \operatorname{tr}\left(\exp\left(\frac{\lambda^2 \sum_{i=1}^n \operatorname{var}(\mathbf{Q}_i)}{1-b|\lambda|}\right)\right) \leq \operatorname{rank}\left(\sum_{i=1}^n \operatorname{var}(\mathbf{Q}_i)\right) e^{\frac{n\lambda^2\sigma^2}{1-b|\lambda|}},$$

where the final inequality follows from the same argument as the proof of Theorem 6.15. Combined with the upper bound (6.37), we find that

$$\mathbb{P}\left[\left|\!\left|\!\left|\frac{1}{n}\sum_{i=1}^n \mathbf{Q}_i\right|\!\right|\!\right|_2 \geq \delta\right] \leq 2 \operatorname{rank}\left(\sum_{i=1}^n \operatorname{var}(\mathbf{Q}_i)\right) e^{\frac{n\sigma^2\lambda^2}{1-b|\lambda|} - \lambda n \delta},$$

valid for all $\lambda \in [0, 1/b)$. Setting $\lambda = \frac{\delta}{\sigma^2 + b\delta} \in (0, \frac{1}{b})$ and simplifying yields the claim (6.42). □

Remarks: Note that the tail bound (6.42) is of the sub-exponential type, with two regimes of behavior depending on the relative sizes of the parameters σ^2 and b. Thus, it is a natural generalization of the classical Bernstein bound for scalar random variables. As with Theorem 6.15, Theorem 6.17 can also be generalized to non-symmetric (and potentially non-square) matrices $\{\mathbf{A}_i\}_{i=1}^n$ by introducing the sequence of $\{\mathbf{Q}_i\}_{i=1}^n$ symmetric matrices defined in equation (6.39), and imposing the Bernstein condition on it. As one special case, if $|\!|\!|\mathbf{A}_i|\!|\!|_2 \leq b$ almost surely, then it can be verified that the matrices $\{\mathbf{Q}_i\}_{i=1}^n$ satisfy the Bernstein condition with b and the quantity

$$\sigma^2 := \max\left\{\left|\!\left|\!\left|\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\mathbf{A}_i\mathbf{A}_i^{\mathrm{T}}]\right|\!\right|\!\right|_2, \left|\!\left|\!\left|\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\mathbf{A}_i^{\mathrm{T}}\mathbf{A}_i]\right|\!\right|\!\right|_2\right\}. \tag{6.43}$$

We provide an instance of this type of transformation in Example 6.18 to follow.

The problem of matrix completion provides an interesting class of examples in which Theorem 6.17 can be fruitfully applied. See Chapter 10 for a detailed description of the underlying problem, which motivates the following discussion.

Example 6.18 (Tail bounds in matrix completion) Consider an i.i.d. sequence of matrices of the form $\mathbf{A}_i = \xi_i \mathbf{X}_i \in \mathbb{R}^{d\times d}$, where ξ_i is a zero-mean sub-exponential variable that satisfies the Bernstein condition with parameter b and variance ν^2, and $\mathbf{X}_i$ is a random "mask matrix", independent from ξ_i, with a single entry equal to d in a position chosen uniformly at random from all d^2 entries, and all remaining entries equal to zero. By construction, for any fixed matrix $\mathbf{\Theta} \in \mathbb{R}^{d\times d}$, we have $\mathbb{E}[\langle\!\langle \mathbf{A}_i, \mathbf{\Theta}\rangle\!\rangle^2] = \nu^2 |\!|\!|\mathbf{\Theta}|\!|\!|_{\mathrm{F}}^2$—a property that plays an important role in our later analysis of matrix completion.

As noted in Example 6.14, apart from constant factors, there is no loss of generality in assuming that the random matrices $\mathbf{A}_i$ have distributions that are symmetric around zero; in this particular, this symmetry condition is equivalent to requiring that the scalar random variables ξ_i and $-\xi_i$ follow the same distribution. Moreover, as defined, the matrices $\mathbf{A}_i$ are not symmetric (meaning that $\mathbf{A}_i \neq \mathbf{A}_i^{\mathrm{T}}$), but as discussed following Theorem 6.17, we can

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press

bound the operator norm $|||\frac{1}{n}\sum_{i=1}^n \mathbf{A}_i|||_2$ in terms of the operator norm $|||\frac{1}{n}\sum_{i=1}^n \mathbf{Q}_i|||_2$, where the symmetrized version $\mathbf{Q}_i \in \mathbb{R}^{2d\times 2d}$ was defined in equation (6.39).

By the independence between ξ_i and $\mathbf{A}_i$ and the symmetric distribution of ξ_i, we have $\mathbb{E}[\mathbf{Q}_i^{2m+1}] = 0$ for all $m = 0, 1, 2, \ldots$. Turning to the even moments, suppose that entry (a, b) is the only non-zero in the mask matrix $\mathbf{X}_i$. We then have

$$\mathbf{Q}_i^{2m} = (\xi_i)^{2m} d^{2m} \begin{bmatrix} \mathbf{D}_a & 0 \\ 0 & \mathbf{D}_b \end{bmatrix} \quad \text{for all } m = 1, 2, \ldots, \tag{6.44}$$

where $\mathbf{D}_a \in \mathbb{R}^{d\times d}$ is the diagonal matrix with a single 1 in entry (a, a), with $\mathbf{D}_b$ defined analogously. By the Bernstein condition, we have $\mathbb{E}[\xi_i^{2m}] \leq \frac{1}{2}(2m)!b^{2m-2}\nu^2$ for all $m = 1, 2, \ldots$.

On the other hand, $\mathbb{E}[\mathbf{D}_a] = \frac{1}{d}\mathbf{I}_d$ since the probability of choosing a in the first coordinate is $1/d$. We thus see that $\operatorname{var}(\mathbf{Q}_i) = \nu^2 d\mathbf{I}_{2d}$. Putting together the pieces, we have shown that

$$\mathbb{E}[\mathbf{Q}_i^{2m}] \leq \frac{1}{2}(2m)!b^{2m-2}\nu^2 d^{2m}\frac{1}{d}\mathbf{I}_{2d} = \frac{1}{2}(2m)!(bd)^{2m-2}\operatorname{var}(\mathbf{Q}_i),$$

showing that $\mathbf{Q}_i$ satisfies the Bernstein condition with parameters bd and

$$\sigma^2 := \left|\left|\left|\frac{1}{n}\sum_{i=1}^n \operatorname{var}(\mathbf{Q}_i)\right|\right|\right|_2 \leq \nu^2 d.$$

Consequently, Theorem 6.17 implies that

$$\mathbb{P}\left[\left|\left|\left|\frac{1}{n}\sum_{i=1}^n \mathbf{A}_i\right|\right|\right|_2 \geq \delta\right] \leq 4de^{-\frac{n\delta^2}{2d(\nu^2+b\delta)}}. \tag{6.45}$$

♣

In certain cases, it is possible to sharpen the dimension dependence of Theorem 6.17—in particular, by replacing the rank-based pre-factor, which can be as large as d, by a quantity that is potentially much smaller. We illustrate one instance of such a sharpened result in the following example.

Example 6.19 (Bernstein bounds with sharpened dimension dependence) Consider a sequence of independent zero-mean random matrices $\mathbf{Q}_i$ bounded as $|||\mathbf{Q}_i|||_2 \leq 1$ almost surely, and suppose that our goal is to upper bound the maximum eigenvalue $\gamma_{\max}(\mathbf{S}_n)$ of the sum $\mathbf{S}_n := \sum_{i=1}^n \mathbf{Q}_i$. Defining the function $\phi(\lambda) := e^\lambda - \lambda - 1$, we note that it is monotonically increasing on the positive real line. Consequently, as verified in Exercise 6.12, for any pair $\delta > 0$, we have

$$\mathbb{P}[\gamma_{\max}(\mathbf{S}_n) \geq \delta] \leq \inf_{\lambda>0} \frac{\operatorname{tr}(\mathbb{E}[\phi(\lambda\mathbf{S}_n)])}{\phi(\lambda\delta)}. \tag{6.46}$$

Moreover, using the fact that $|||\mathbf{Q}_i|||_2 \leq 1$, the same exercise shows that

$$\log \Psi_{\mathbf{Q}_i}(\lambda) \leq \phi(\lambda)\operatorname{var}(\mathbf{Q}_i) \tag{6.47a}$$

and

$$\operatorname{tr}(\mathbb{E}[\phi(\lambda\mathbf{S}_n)]) \leq \frac{\operatorname{tr}(\bar{\mathbf{V}})}{|||\bar{\mathbf{V}}|||_2} e^{\phi(\lambda)|||\bar{\mathbf{V}}|||_2}, \tag{6.47b}$$

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press

where $\bar{\mathbf{V}} := \sum_{i=1}^{n} \mathrm{var}(\mathbf{Q}_i)$. Combined with the initial bound (6.46), we conclude that

$$\mathbb{P}[\gamma_{\max}(\mathbf{S}_n) \geq \delta] \leq \frac{\mathrm{tr}(\bar{\mathbf{V}})}{|||\bar{\mathbf{V}}|||_2} \inf_{\lambda>0} \left\{ \frac{e^{\phi(\lambda)|||\bar{\mathbf{V}}|||_2}}{\phi(\lambda\delta)} \right\}. \tag{6.48}$$

The significance of this bound is the appearance of the trace ratio $\frac{\mathrm{tr}(\bar{\mathbf{V}})}{|||\bar{\mathbf{V}}|||_2}$ as a pre-factor, as opposed to the quantity $\mathrm{rank}(\bar{\mathbf{V}}) \leq d$ that arose in our previous derivation. Note that we always have $\frac{\mathrm{tr}(\bar{\mathbf{V}})}{|||\bar{\mathbf{V}}|||_2} \leq \mathrm{rank}(\bar{\mathbf{V}})$, and in certain cases, the trace ratio can be substantially smaller than the rank. See Exercise 6.13 for one such case. ♣

6.4.5 Consequences for covariance matrices

We conclude with a useful corollary of Theorem 6.17 for the estimation of covariance matrices.

Corollary 6.20 *Let $x_1, \ldots, x_n$ be i.i.d. zero-mean random vectors with covariance $\mathbf{\Sigma}$ such that $\|x_j\|_2 \leq \sqrt{b}$ almost surely. Then for all $\delta > 0$, the sample covariance matrix $\widehat{\mathbf{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^{\mathrm{T}}$ satisfies*

$$\mathbb{P}[|||\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}|||_2 \geq \delta] \leq 2d \exp\left(-\frac{n\delta^2}{2b(|||\mathbf{\Sigma}|||_2 + \delta)} \right). \tag{6.49}$$

Proof We apply Theorem 6.17 to the zero-mean random matrices $\mathbf{Q}_i := x_i x_i^{\mathrm{T}} - \mathbf{\Sigma}$. These matrices have controlled operator norm: indeed, by the triangle inequality, we have

$$|||\mathbf{Q}_i|||_2 \leq \|x_i\|_2^2 + |||\mathbf{\Sigma}|||_2 \leq b + |||\mathbf{\Sigma}|||_2.$$

Since $\mathbf{\Sigma} = \mathbb{E}[x_i x_i^{\mathrm{T}}]$, we have $|||\mathbf{\Sigma}|||_2 = \max_{v \in \mathbb{S}^{d-1}} \mathbb{E}[\langle v,\, x_i \rangle^2] \leq b$, and hence $|||\mathbf{Q}_i|||_2 \leq 2b$. Turning to the variance of $\mathbf{Q}_i$, we have

$$\mathrm{var}(\mathbf{Q}_i) = \mathbb{E}[(x_i x_i^{\mathrm{T}})^2] - \mathbf{\Sigma}^2 \preceq \mathbb{E}[\|x_i\|_2^2\, x_i x_i^{\mathrm{T}}] \preceq b\mathbf{\Sigma},$$

so that $|||\,\mathrm{var}(\mathbf{Q}_i)|||_2 \leq b|||\mathbf{\Sigma}|||_2$. Substituting into the tail bound (6.42) yields the claim. □

Let us illustrate some consequences of this corollary with some examples.

Example 6.21 (Random vectors uniform on a sphere) Suppose that the random vectors x_i are chosen uniformly from the sphere $\mathbb{S}^{d-1}(\sqrt{d})$, so that $\|x_i\|_2 = \sqrt{d}$ for all $i = 1, \ldots, n$. By construction, we have $\mathbb{E}[x_i x_i^{\mathrm{T}}] = \mathbf{\Sigma} = \mathbf{I}_d$, and hence $|||\mathbf{\Sigma}|||_2 = 1$. Applying Corollary 6.20 yields

$$\mathbb{P}[|||\widehat{\mathbf{\Sigma}} - \mathbf{I}_d|||_2 \geq \delta] \leq 2de^{-\frac{n\delta^2}{2d+2d\delta}} \qquad \text{for all } \delta \geq 0. \tag{6.50}$$

This bound implies that

$$|||\widehat{\mathbf{\Sigma}} - \mathbf{I}_d|||_2 \precsim \sqrt{\frac{d \log d}{n}} + \frac{d \log d}{n} \tag{6.51}$$

with high probability, so that the sample covariance is a consistent estimate as long as $\frac{d \log d}{n} \to 0$. This result is close to optimal, with only the extra logarithmic factor being superfluous in this particular case. It can be removed, for instance, by noting that x_i is a sub-Gaussian random vector, and then applying Theorem 6.5. ♣

Example 6.22 ("Spiked" random vectors) We now consider an ensemble of random vectors that are rather different than the previous example, but still satisfy the same bound. In particular, consider a random vector of the form $x_i = \sqrt{d} e_{a(i)}$, where $a(i)$ is an index chosen uniformly at random from $\{1, \ldots, d\}$, and $e_{a(i)} \in \mathbb{R}^d$ is the canonical basis vector with 1 in position $a(i)$. As before, we have $\|x_i\|_2 = \sqrt{d}$, and $\mathbb{E}[x_i x_i^{\mathrm{T}}] = \mathbf{I}_d$ so that the tail bound (6.50) also applies to this ensemble. An interesting fact is that, for this particular ensemble, the bound (6.51) is sharp, meaning it cannot be improved beyond constant factors. ♣

6.5 Bounds for structured covariance matrices

In the preceding sections, our primary focus has been estimation of general unstructured covariance matrices via the sample covariance. When a covariance matrix is equipped with additional structure, faster rates of estimation are possible using different estimators than the sample covariance matrix. In this section, we explore the faster rates that are achievable for sparse and/or graph-structured matrices.

In the simplest setting, the covariance matrix is known to be sparse, and the positions of the non-zero entries are known. In such settings, it is natural to consider matrix estimators that are non-zero only in these known positions. For instance, if we are given *a priori* knowledge that the covariance matrix is diagonal, then it would be natural to use the estimate $\widehat{\mathbf{D}} := \operatorname{diag}\{\widehat{\Sigma}_{11}, \widehat{\Sigma}_{22}, \ldots, \widehat{\Sigma}_{dd}\}$, corresponding to the diagonal entries of the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$. As we explore in Exercise 6.15, the performance of this estimator can be substantially better: in particular, for sub-Gaussian variables, it achieves an estimation error of the order $\sqrt{\frac{\log d}{n}}$, as opposed to the order $\sqrt{\frac{d}{n}}$ rates in the unstructured setting. Similar statements apply to other forms of known sparsity.

6.5.1 Unknown sparsity and thresholding

More generally, suppose that the covariance matrix $\boldsymbol{\Sigma}$ is known to be relatively sparse, but that the positions of the non-zero entries are no longer known. It is then natural to consider estimators based on thresholding. Given a parameter $\lambda > 0$, the *hard-thresholding operator* is given by

$$T_\lambda(u) := u\, \mathbb{I}[|u| > \lambda] = \begin{cases} u & \text{if } |u| > \lambda, \\ 0 & \text{otherwise.} \end{cases} \tag{6.52}$$

With a minor abuse of notation, for a matrix $\mathbf{M}$, we write $T_\lambda(\mathbf{M})$ for the matrix obtained by applying the thresholding operator to each element of $\mathbf{M}$. In this section, we study the performance of the estimator $T_{\lambda_n}(\widehat{\boldsymbol{\Sigma}})$, where the parameter $\lambda_n > 0$ is suitably chosen as a function of the sample size n and matrix dimension d.

The sparsity of the covariance matrix can be measured in various ways. Its zero pattern

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press

is captured by the adjacency matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ with entries $A_{j\ell} = \mathbb{I}[\Sigma_{j\ell} \neq 0]$. This adjacency matrix defines the edge structure of an undirected graph G on the vertices $\{1, 2, \ldots, d\}$, with edge (j, ℓ) included in the graph if and only if $\Sigma_{j\ell} \neq 0$, along with the self-edges (j, j) for each of the diagonal entries. The operator norm $|||\mathbf{A}|||_2$ of the adjacency matrix provides a natural measure of sparsity. In particular, it can be verified that $|||\mathbf{A}|||_2 \leq d$, with equality holding when G is fully connected, meaning that $\mathbf{\Sigma}$ has no zero entries. More generally, as shown in Exercise 6.2, we have $|||\mathbf{A}|||_2 \leq s$ whenever $\mathbf{\Sigma}$ has at most s non-zero entries per row, or equivalently when the graph G has maximum degree at most $s - 1$. The following result provides a guarantee for the thresholded sample covariance matrix that involves the graph adjacency matrix $\mathbf{A}$ defined by $\mathbf{\Sigma}$.

Theorem 6.23 (Thresholding-based covariance estimation) *Let $\{x_i\}_{i=1}^n$ be an i.i.d. sequence of zero-mean random vectors with covariance matrix $\mathbf{\Sigma}$, and suppose that each component x_{ij} is sub-Gaussian with parameter at most σ. If $n > \log d$, then for any $\delta > 0$, the thresholded sample covariance matrix $T_{\lambda_n}(\widehat{\mathbf{\Sigma}})$ with $\lambda_n/\sigma^2 = 8\sqrt{\frac{\log d}{n}} + \delta$ satisfies*

$$\mathbb{P}[|||T_{\lambda_n}(\widehat{\mathbf{\Sigma}}) - \mathbf{\Sigma}|||_2 \geq 2|||\mathbf{A}|||_2\lambda_n] \leq 8e^{-\frac{n}{16}\min\{\delta,\delta^2\}}. \tag{6.53}$$

Underlying the proof of Theorem 6.23 is the following (deterministic) result: for any choice of λ_n such that $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max} \leq \lambda_n$, we are guaranteed that

$$|||T_{\lambda_n}(\widehat{\mathbf{\Sigma}}) - \mathbf{\Sigma}|||_2 \leq 2|||\mathbf{A}|||_2\lambda_n. \tag{6.54}$$

The proof of this intermediate claim is straightforward. First, for any index pair (j, ℓ) such that $\Sigma_{j\ell} = 0$, the bound $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max} \leq \lambda_n$ guarantees that $|\widehat{\Sigma}_{j\ell}| \leq \lambda_n$, and hence that $T_{\lambda_n}(\widehat{\Sigma}_{j\ell}) = 0$ by definition of the thresholding operator. On the other hand, for any pair (j, ℓ) for which $\Sigma_{j\ell} \neq 0$, we have

$$|T_{\lambda_n}(\widehat{\Sigma}_{j\ell}) - \Sigma_{j\ell}| \overset{\text{(i)}}{\leq} |T_{\lambda_n}(\widehat{\Sigma}_{j\ell}) - \widehat{\Sigma}_{j\ell}| + |\widehat{\Sigma}_{j\ell} - \Sigma_{j\ell}| \overset{\text{(ii)}}{\leq} 2\lambda_n,$$

where step (i) follows from the triangle inequality, and step (ii) follows from the fact that $|T_{\lambda_n}(\widehat{\Sigma}_{j\ell}) - \widehat{\Sigma}_{j\ell}| \leq \lambda_n$, and a second application of the assumption $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max} \leq \lambda_n$. Consequently, we have shown that the matrix $\mathbf{B} := |T_{\lambda_n}(\widehat{\mathbf{\Sigma}}) - \mathbf{\Sigma}|$ satisfies the elementwise inequality $\mathbf{B} \leq 2\lambda_n\mathbf{A}$. Since both $\mathbf{B}$ and $\mathbf{A}$ have non-negative entries, we are guaranteed that $|||\mathbf{B}|||_2 \leq 2\lambda_n|||\mathbf{A}|||_2$, and hence that $|||T_{\lambda_n}(\widehat{\mathbf{\Sigma}}) - \mathbf{\Sigma}|||_2 \leq 2\lambda_n|||\mathbf{A}|||_2$ as claimed. (See Exercise 6.3 for the details of these last steps.)

Theorem 6.23 has a number of interesting corollaries for particular classes of covariance matrices.

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press

Corollary 6.24 *Suppose that, in addition to the conditions of Theorem 6.23, the covariance matrix $\boldsymbol{\Sigma}$ has at most s non-zero entries per row. Then with $\lambda_n/\sigma^2 = 8\sqrt{\frac{\log d}{n}} + \delta$ for some $\delta > 0$, we have*

$$\mathbb{P}[|||T_{\lambda_n}(\widehat{\boldsymbol{\Sigma}}) - \boldsymbol{\Sigma}|||_2 \geq 2s\lambda_n] \leq 8e^{-\frac{n}{16}\min\{\delta,\delta^2\}}. \tag{6.55}$$

In order to establish these claims from Theorem 6.23, it suffices to show that $|||\mathbf{A}|||_2 \leq s$. Since $\mathbf{A}$ has at most s ones per row (with the remaining entries equal to zero), this claim follows from the result of Exercise 6.2.

Example 6.25 (Sparsity and adjacency matrices) In certain ways, the bound (6.55) is more appealing than the bound (6.53), since it is based on a local quantity—namely, the maximum degree of the graph defined by the covariance matrix, as opposed to the spectral norm $|||\mathbf{A}|||_2$. In certain cases, these two bounds coincide. As an example, consider any graph with maximum degree $s - 1$ that contains an s-clique (i.e., a subset of s nodes that are all joined by edges). As we explore in Exercise 6.16, for any such graph, we have $|||\mathbf{A}|||_2 = s$, so that the two bounds are equivalent.

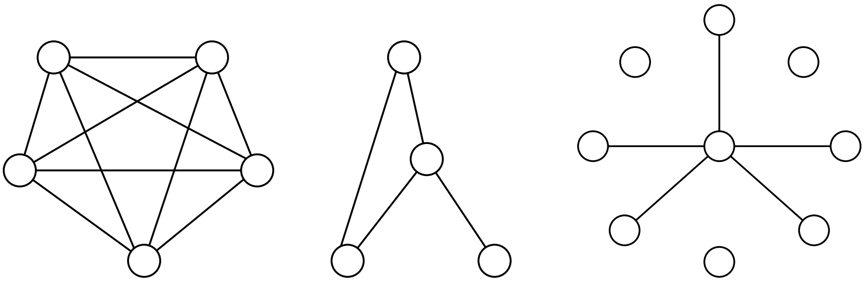

Figure 6.1 (a) An instance of a graph on $d = 9$ nodes containing an $s = 5$ clique. For this class of graphs, the bounds (6.53) and (6.55) coincide. (b) A hub-and-spoke graph on $d = 9$ nodes with maximum degree $s = 5$. For this class of graphs, the bounds differ by a factor of $\sqrt{s}$.

However, in general, the bound (6.53) can be substantially sharper than the bound (6.55). As an example, consider a hub-and-spoke graph, in which one central node known as the hub is connected to s of the remaining $d - 1$ nodes, as illustrated in Figure 6.1(b). For such a graph, we have $|||\mathbf{A}|||_2 = 1 + \sqrt{s-1}$, so that in this case Theorem 6.23 guarantees that

$$|||T_{\lambda_n}(\widehat{\boldsymbol{\Sigma}}) - \boldsymbol{\Sigma}|||_2 \precsim \sqrt{\frac{s \log d}{n}},$$

with high probability, a bound that is sharper by a factor of order $\sqrt{s}$ compared to the bound (6.55) from Corollary 6.24. ♣

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press

We now turn to the proof of the remainder of Theorem 6.23. Based on the reasoning leading to equation (6.54), it suffices to establish a high-probability bound on the elementwise infinity norm of the error matrix $\widehat{\mathbf{\Delta}} := \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}$.

Lemma 6.26 *Under the conditions of Theorem 6.23, we have*

$$\mathbb{P}[\|\widehat{\mathbf{\Delta}}\|_{\max}/\sigma^2 \geq t] \leq 8e^{-\frac{n}{16}\min\{t,\, t^2\}+2\log d} \qquad \textit{for all } t > 0. \tag{6.56}$$

Setting $t = \lambda_n/\sigma^2 = 8\sqrt{\frac{\log d}{n}} + \delta$ in the bound (6.56) yields

$$\mathbb{P}[\|\widehat{\mathbf{\Delta}}\|_{\max} \geq \lambda_n] \leq 8e^{-\frac{n}{16}\min\{\delta,\, \delta^2\}},$$

where we have used the fact that $n > \log d$ by assumption.

It remains to prove Lemma 6.26. Note that the rescaled vector x_i/σ is sub-Gaussian with parameter at most 1. Consequently, we may assume without loss of generality that $\sigma = 1$, and then rescale at the end. First considering a diagonal entry, the result of Exercise 6.15(a) guarantees that there are universal positive constants c_1, c_2 such that

$$\mathbb{P}[|\widehat{\Delta}_{jj}| \geq c_1\delta] \leq 2e^{-c_2 n\delta^2} \qquad \text{for all } \delta \in (0,1). \tag{6.57}$$

Turning to the non-diagonal entries, for any $j \neq \ell$, we have

$$2\widehat{\Delta}_{j\ell} = \frac{2}{n}\sum_{i=1}^{n} x_{ij}x_{i\ell} - 2\Sigma_{j\ell} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij}+x_{i\ell})^2 - (\Sigma_{jj}+\Sigma_{\ell\ell}+2\Sigma_{j\ell}) - \widehat{\Delta}_{jj} - \widehat{\Delta}_{\ell\ell}.$$

Since x_{ij} and $x_{i\ell}$ are both zero-mean and sub-Gaussian with parameter σ, the sum $x_{ij} + x_{i\ell}$ is zero-mean and sub-Gaussian with parameter at most $2\sqrt{2}\sigma$ (see Exercise 2.13(c)). Consequently, there are universal constants c_2, c_3 such that for all $\delta \in (0,1)$, we have

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n}(x_{ij}+x_{i\ell})^2 - (\Sigma_{jj}+\Sigma_{\ell\ell}+2\Sigma_{j\ell})\right| \geq c_3\delta\right] \leq 2e^{-c_2 n\delta^2},$$

and hence, combining with our earlier diagonal bound (6.57), we obtain the tail bound $\mathbb{P}[|\widehat{\Delta}_{j\ell}| \geq c_1'\delta] \leq 6e^{-c_2 n\delta^2}$. Finally, combining this bound with the earlier inequality (6.57) and then taking a union bound over all d^2 entries of the matrix yields the stated claim (6.56).

6.5.2 Approximate sparsity

Given a covariance matrix $\mathbf{\Sigma}$ with no entries that are exactly zero, the bounds of Theorem 6.23 are very poor. In particular, for a completely dense matrix, the associated adjacency matrix $\mathbf{A}$ is simply the all-ones matrix, so that $|||\mathbf{A}|||_2 = d$. Intuitively, one might expect that these bounds could be improved if $\mathbf{\Sigma}$ had a large number of non-zero entries, but many of them were "near zero".

Recall that one way in which to measure the sparsity of $\mathbf{\Sigma}$ is in terms of the maximum number of non-zero entries per row. A generalization of this idea is to measure the ℓ_q-"norm"

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press

of each row. More specifically, given a parameter $q \in [0, 1]$ and a radius R_q, we impose the constraint

$$\max_{j=1,\ldots,d} \sum_{\ell=1}^{d} |\Sigma_{j\ell}|^q \leq R_q. \tag{6.58}$$

(See Figure 7.1 in Chapter 7 for an illustration of these types of sets.) In the special case $q = 0$, this constraint is equivalent to requiring that each row of $\mathbf{\Sigma}$ have at most R_0 non-zero entries. For intermediate values $q \in (0, 1]$, it allows for many non-zero entries but requires that their absolute magnitudes (if ordered from largest to smallest) drop off relatively quickly.

Theorem 6.27 (Covariance estimation under ℓ_q-sparsity) *Suppose that the covariance matrix* $\mathbf{\Sigma}$ *satisfies the* ℓ_q*-sparsity constraint* (6.58). *Then for any* λ_n *such that* $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max} \leq \lambda_n/2$, *we are guaranteed that*

$$|||T_{\lambda_n}(\widehat{\mathbf{\Sigma}}) - \mathbf{\Sigma}|||_2 \leq 4R_q{\lambda_n}^{1-q}. \tag{6.59a}$$

Consequently, if the sample covariance is formed using i.i.d. samples $\{x_i\}_{i=1}^n$ *that are zero-mean with sub-Gaussian parameter at most* σ, *then with* $\lambda_n/\sigma^2 = 8\sqrt{\frac{\log d}{n}} + \delta$, *we have*

$$\mathbb{P}[|||T_{\lambda_n}(\widehat{\mathbf{\Sigma}}) - \mathbf{\Sigma}|||_2 \geq 4R_q{\lambda_n}^{1-q}] \leq 8e^{-\frac{n}{16}\min\{\delta, \delta^2\}} \quad \textit{for all } \delta > 0. \tag{6.59b}$$

Proof Given the deterministic claim (6.59a), the probabilistic bound (6.59b) follows from standard tail bounds on sub-exponential variables. The deterministic claim is based on the assumption that $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max} \leq \lambda_n/2$. By the result of Exercise 6.2, the operator norm can be upper bounded as

$$|||T_{\lambda_n}(\widehat{\mathbf{\Sigma}}) - \mathbf{\Sigma}|||_2 \leq \max_{j=1,\ldots,d} \sum_{\ell=1}^{d} |T_{\lambda_n}(\widehat{\Sigma}_{j\ell}) - \Sigma_{j\ell}|.$$

Fixing an index $j \in \{1, 2, \ldots, d\}$, define the set $S_j(\lambda_n/2) = \Big\{\ell \in \{1, \ldots, d\} \mid |\Sigma_{j\ell}| > \lambda_n/2\Big\}$. For any index $\ell \in S_j(\lambda_n/2)$, we have

$$|T_{\lambda_n}(\widehat{\Sigma}_{j\ell}) - \Sigma_{j\ell}| \leq |T_{\lambda_n}(\widehat{\Sigma}_{j\ell}) - \widehat{\Sigma}_{j\ell}| + |\widehat{\Sigma}_{j\ell} - \Sigma_{j\ell}| \leq \frac{3}{2}\lambda_n.$$

On the other hand, for any index $\ell \notin S_j(\lambda_n/2)$, we have $T_{\lambda_n}(\widehat{\Sigma}_{j\ell}) = 0$, by definition of the thresholding operator, and hence

$$|T_{\lambda_n}(\widehat{\Sigma}_{j\ell}) - \Sigma_{j\ell}| = |\Sigma_{j\ell}|.$$

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press

Putting together the pieces, we have

$$\begin{aligned}\sum_{\ell=1}^{d} |T_{\lambda_n}(\widehat{\Sigma}_{j\ell}) - \Sigma_{j\ell}| &= \sum_{\ell \in S_j(\lambda_n)} |T_{\lambda_n}(\widehat{\Sigma}_{j\ell}) - \Sigma_{j\ell}| + \sum_{\ell \notin S_j(\lambda_n)} |T_{\lambda_n}(\widehat{\Sigma}_{j\ell}) - \Sigma_{j\ell}| \\ &\leq |S_j(\lambda_n/2)| \frac{3}{2} \lambda_n + \sum_{\ell \notin S_j(\lambda_n)} |\Sigma_{j\ell}|. \end{aligned} \tag{6.60}$$

Now we have

$$\sum_{\ell \notin S_j(\lambda_n/2)} |\Sigma_{j\ell}| = \frac{\lambda_n}{2} \sum_{\ell \notin S_j(\lambda_n/2)} \frac{|\Sigma_{j\ell}|}{\lambda_n/2} \overset{\text{(i)}}{\leq} \frac{\lambda_n}{2} \sum_{\ell \notin S_j(\lambda_n/2)} \left(\frac{|\Sigma_{j\ell}|}{\lambda_n/2} \right)^q \overset{\text{(ii)}}{\leq} \lambda_n{}^{1-q} R_q,$$

where step (i) follows since $|\Sigma_{j\ell}| \leq \lambda_n/2$ for all $\ell \notin S_j(\lambda_n/2)$ and $q \in [0, 1]$, and step (ii) follows by the assumption (6.58). On the other hand, we have

$$R_q \geq \sum_{\ell=1}^{d} |\Sigma_{j\ell}|^q \geq |S_j(\lambda_n/2)| \left(\frac{\lambda_n}{2} \right)^q,$$

whence $|S_j(\lambda_n/2)| \leq 2^q R_q \lambda_n^{-q}$. Combining these ingredients with the inequality (6.60), we find that

$$\sum_{\ell=1}^{d} |T_{\lambda_n}(\widehat{\Sigma}_{j\ell}) - \Sigma_{j\ell}| \leq 2^q R_q \lambda_n^{1-q} \frac{3}{2} + R_q \lambda_n^{1-q} \leq 4 R_q \lambda_n^{1-q}.$$

Since this same argument holds for each index $j = 1, \ldots, d$, the claim (6.59a) follows. □

6.6 Appendix: Proof of Theorem 6.1

It remains to prove the lower bound (6.9) on the minimal singular value. In order to do so, we follow an argument similar to that used to upper bound the maximal singular value. Throughout this proof, we assume that $\mathbf{\Sigma}$ is strictly positive definite (and hence invertible); otherwise, its minimal singular value is zero, and the claimed lower bound is vacuous. We begin by lower bounding the expectation using a Gaussian comparison principle due to Gordon (1985). By definition, the minimum singular value has the variational representation $\sigma_{\min}(\mathbf{X}) = \min_{v' \in \mathbb{S}^{d-1}} \|\mathbf{X}v'\|_2$. Let us reformulate this representation slightly for later theoretical convenience. Recalling the shorthand notation $\bar{\sigma}_{\min} = \sigma_{\min}(\sqrt{\mathbf{\Sigma}})$, we define the radius $R = 1/\bar{\sigma}_{\min}$, and then consider the set

$$\mathcal{V}(R) := \{z \in \mathbb{R}^d \mid \|\sqrt{\mathbf{\Sigma}}z\|_2 = 1,\ \|z\|_2 \leq R\}. \tag{6.61}$$

We claim that it suffices to show that a lower bound of the form

$$\min_{z \in \mathcal{V}(R)} \frac{\|\mathbf{X}z\|_2}{\sqrt{n}} \geq 1 - \delta - R\sqrt{\frac{\operatorname{tr}(\mathbf{\Sigma})}{n}} \tag{6.62}$$

holds with probability at least $1 - e^{-n\delta^2/2}$. Indeed, suppose that inequality (6.62) holds. Then for any $v' \in \mathbb{S}^{d-1}$, we can define the rescaled vector $z := \frac{v'}{\|\sqrt{\mathbf{\Sigma}}v'\|_2}$. By construction, we have

$$\|\sqrt{\mathbf{\Sigma}}z\|_2 = 1 \quad \text{and} \quad \|z\|_2 = \frac{1}{\|\sqrt{\mathbf{\Sigma}}v'\|_2} \leq \frac{1}{\sigma_{\min}(\sqrt{\mathbf{\Sigma}})} = R,$$

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press

so that $z \in \mathcal{V}(R)$. We now observe that

$$\frac{\|\mathbf{X}v'\|_2}{\sqrt{n}} = \|\sqrt{\mathbf{\Sigma}}v'\|_2 \frac{\|\mathbf{X}z\|_2}{\sqrt{n}} \geq \sigma_{\min}(\sqrt{\mathbf{\Sigma}}) \min_{z \in \mathcal{V}(R)} \frac{\|\mathbf{X}z\|_2}{\sqrt{n}}.$$

Since this bound holds for all $v' \in \mathbb{S}^{d-1}$, we can take the minimum on the left-hand side, thereby obtaining

$$\begin{aligned} \min_{v' \in \mathbb{S}^{d-1}} \frac{\|\mathbf{X}v'\|_2}{\sqrt{n}} &\geq \bar{\sigma}_{\min} \min_{z \in \mathcal{V}(R)} \frac{\|\mathbf{X}z\|_2}{\sqrt{n}} \\ &\overset{\text{(i)}}{\geq} \bar{\sigma}_{\min} \left\{ 1 - R\sqrt{\frac{\operatorname{tr}(\mathbf{\Sigma})}{n}} - \delta \right\} \\ &= (1-\delta)\bar{\sigma}_{\min} - R\sqrt{\frac{\operatorname{tr}(\mathbf{\Sigma})}{n}}, \end{aligned}$$

where step (i) follows from the bound (6.62).

It remains to prove the lower bound (6.62). We begin by showing concentration of the random variable $\min_{v \in \mathcal{V}(R)} \|\mathbf{X}v\|_2/\sqrt{n}$ around its expected value. Since the matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ has i.i.d. rows, each drawn from the $\mathcal{N}(0, \mathbf{\Sigma})$ distribution, we can write $\mathbf{X} = \mathbf{W}\sqrt{\mathbf{\Sigma}}$, where the random matrix $\mathbf{W}$ is standard Gaussian. Using the fact that $\|\sqrt{\mathbf{\Sigma}}v\|_2 = 1$ for all $v \in \mathcal{V}(R)$, it follows that the function $\mathbf{W} \mapsto \min_{v \in \mathcal{V}(R)} \frac{\|\mathbf{W}\sqrt{\mathbf{\Sigma}}v\|_2}{\sqrt{n}}$ is Lipschitz with parameter $L = 1/\sqrt{n}$. Applying Theorem 2.26, we conclude that

$$\min_{v \in \mathcal{V}(R)} \frac{\|\mathbf{X}v\|_2}{\sqrt{n}} \geq \mathbb{E}\left[\min_{v \in \mathcal{V}(R)} \frac{\|\mathbf{X}v\|_2}{\sqrt{n}} \right] - \delta$$

with probability at least $1 - e^{-n\delta^2/2}$.

Consequently, the proof will be complete if we can show that

$$\mathbb{E}\left[\min_{v \in \mathcal{V}(R)} \frac{\|\mathbf{X}v\|_2}{\sqrt{n}} \right] \geq 1 - R\sqrt{\frac{\operatorname{tr}(\mathbf{\Sigma})}{n}}. \tag{6.63}$$

In order to do so, we make use of an extension of the Sudakov–Fernique inequality, known as Gordon's inequality, which we now state. Let $\{Z_{u,v}\}$ and $\{Y_{u,v}\}$ be a pair of zero-mean Gaussian processes indexed by a non-empty index set $\mathbb{T} = U \times V$. Suppose that

$$\mathbb{E}[(Z_{u,v} - Z_{\widetilde{u},\widetilde{v}})^2] \leq \mathbb{E}[(Y_{u,v} - Y_{\widetilde{u},\widetilde{v}})^2] \qquad \text{for all pairs } (u,v) \text{ and } (\widetilde{u},\widetilde{v}) \in \mathbb{T}, \tag{6.64}$$

and moreover that this inequality holds with *equality* whenever $v = \widetilde{v}$. Under these conditions, Gordon's inequality guarantees that

$$\mathbb{E}\left[\max_{v \in V} \min_{u \in U} Z_{u,v} \right] \leq \mathbb{E}\left[\max_{v \in V} \min_{u \in U} Y_{u,v} \right]. \tag{6.65}$$

In order to exploit this result, we first observe that

$$-\min_{z \in \mathcal{V}(R)} \|\mathbf{X}z\|_2 = \max_{z \in \mathcal{V}(R)} \{-\|\mathbf{X}z\|_2\} = \max_{z \in \mathcal{V}(R)} \min_{u \in \mathbb{S}^{n-1}} u^{\mathrm{T}}\mathbf{X}z.$$

As before, if we introduce the standard Gaussian random matrix $\mathbf{W} \in \mathbb{R}^{n \times d}$, then for any

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press

$z \in \mathcal{V}(R)$, we can write $u^{\mathrm{T}}\mathbf{X}z = u^{\mathrm{T}}\mathbf{W}v$, where $v := \sqrt{\mathbf{\Sigma}}z$. Whenever $z \in \mathcal{V}(R)$, then the vector v must belong to the set $\mathcal{V}'(R) := \{v \in \mathbb{S}^{d-1} \mid \|\mathbf{\Sigma}^{-\frac{1}{2}}v\|_2 \leq R\}$, and we have shown that

$$\min_{z \in \mathcal{V}(R)} \|\mathbf{X}z\|_2 = \max_{v \in \mathcal{V}'(R)} \min_{u \in \mathbb{S}^{n-1}} \underbrace{u^{\mathrm{T}}\mathbf{W}v}_{Z_{u,v}}.$$

Let (u, v) and $(\widetilde{u}, \widetilde{v})$ be any two members of the Cartesian product space $\mathbb{S}^{n-1} \times \mathcal{V}'(R)$. Since $\|u\|_2 = \|\widetilde{u}\|_2 = \|v\|_2 = \|\widetilde{v}\|_2 = 1$, following the same argument as in bounding the maximal singular value shows that

$$\rho_Z^2((u, v), (\widetilde{u}, \widetilde{v})) \leq \|u - \widetilde{u}\|_2^2 + \|v - \widetilde{v}\|_2^2, \tag{6.66}$$

with equality holding when $v = \widetilde{v}$. Consequently, if we define the Gaussian process $Y_{u,v} := \langle g, u\rangle + \langle h, v\rangle$, where $g \in \mathbb{R}^n$ and $h \in \mathbb{R}^d$ are standard Gaussian vectors and mutually independent, then we have

$$\rho_Y^2((u, v), (\widetilde{u}, \widetilde{v})) = \|u - \widetilde{u}\|_2^2 + \|v - \widetilde{v}\|_2^2,$$

so that the Sudakov–Fernique increment condition (6.64) holds. In addition, for a pair such that $v = \widetilde{v}$, equality holds in the upper bound (6.66), which guarantees that $\rho_Z((u, v), (\widetilde{u}, v)) = \rho_Y((u, v), (\widetilde{u}, v))$. Consequently, we may apply Gordon's inequality (6.65) to conclude that

$$\begin{aligned}
\mathbb{E}\Big[- \min_{z \in \mathcal{V}(R)} \|\mathbf{X}z\|_2 \Big] &\leq \mathbb{E}\Big[\max_{v \in \mathcal{V}'(R)} \min_{u \in \mathbb{S}^{n-1}} Y_{u,v} \Big] \\
&= \mathbb{E}\Big[\min_{u \in \mathbb{S}^{n-1}} \langle g, u\rangle \Big] + \mathbb{E}\Big[\max_{v \in \mathcal{V}'(R)} \langle h, v\rangle \Big] \\
&\leq -\mathbb{E}[\|g\|_2] + \mathbb{E}[\|\sqrt{\mathbf{\Sigma}}h\|_2]R,
\end{aligned}$$

where we have used the upper bound $|\langle h, v\rangle| = |\langle \sqrt{\mathbf{\Sigma}}h, \mathbf{\Sigma}^{-\frac{1}{2}}v\rangle| \leq \|\sqrt{\mathbf{\Sigma}}h\|_2 R$, by definition of the set $\mathcal{V}'(R)$.

We now claim that

$$\frac{\mathbb{E}[\|\sqrt{\mathbf{\Sigma}}h\|_2]}{\sqrt{\mathrm{tr}(\mathbf{\Sigma})}} \leq \frac{\mathbb{E}[\|h\|_2]}{\sqrt{d}}. \tag{6.67}$$

Indeed, by the rotation invariance of the Gaussian distribution, we may assume that $\mathbf{\Sigma}$ is diagonal, with non-negative entries $\{\gamma_j\}_{j=1}^d$, and the claim is equivalent to showing that the function $F(\gamma) := \mathbb{E}[(\sum_{j=1}^d \gamma_j h_j^2)^{1/2}]$ achieves its maximum over the probability simplex at the uniform vector (i.e., with all entries $\gamma_j = 1/d$). Since F is continuous and the probability simplex is compact, the maximum is achieved. By the rotation invariance of the Gaussian, the function F is also permutation invariant—i.e., $F(\gamma) = F(\Pi(\gamma))$ for all permutation matrices Π. Since F is also concave, the maximum must be achieved at $\gamma_j = 1/d$, which establishes the inequality (6.67).

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press

Recalling that $R = 1/\bar{\sigma}_{\min}$, we then have

$$\begin{aligned}-\mathbb{E}[\|g\|_2] + R\,\mathbb{E}[\|\sqrt{\Sigma}h\|_2] &\leq -\mathbb{E}[\|g\|_2] + \frac{\sqrt{\mathrm{tr}(\mathbf{\Sigma})}}{\bar{\sigma}_{\min}}\frac{\mathbb{E}[\|h\|_2]}{\sqrt{d}} \\ &= \underbrace{\{-\mathbb{E}[\|g\|_2] + \mathbb{E}[\|h\|_2]\}}_{T_1} + \underbrace{\left\{\sqrt{\frac{\mathrm{tr}(\mathbf{\Sigma})}{\bar{\sigma}^2_{\min} d}} - 1\right\}\mathbb{E}[\|h\|_2]}_{T_2}.\end{aligned}$$

By Jensen's inequality, we have $\mathbb{E}\|h\|_2 \leq \sqrt{\mathbb{E}\|h\|_2^2} = \sqrt{d}$. Since $\frac{\mathrm{tr}(\Sigma)}{\bar{\sigma}^2_{\min} d} \geq 1$, we conclude that $T_2 \leq \left\{\sqrt{\frac{\mathrm{tr}(\Sigma)}{\bar{\sigma}^2_{\min} d}} - 1\right\}\sqrt{d}$. On the other hand, a direct calculation, using our assumption that $n \geq d$, shows that $T_1 \leq -\sqrt{n} + \sqrt{d}$. Combining the pieces, we conclude that

$$\begin{aligned}\mathbb{E}\left[-\min_{z\in\mathcal{V}(R)}\|\mathbf{X}z\|_2\right] &\leq -\sqrt{n} + \sqrt{d} + \left\{\sqrt{\frac{\mathrm{tr}(\mathbf{\Sigma})}{\bar{\sigma}^2_{\min} d}} - 1\right\}\sqrt{d} \\ &= -\sqrt{n} + \frac{\sqrt{\mathrm{tr}(\mathbf{\Sigma})}}{\bar{\sigma}_{\min}},\end{aligned}$$

which establishes the initial claim (6.62), thereby completing the proof.

6.7 Bibliographic details and background

The two-volume series by Horn and Johnson (1985; 1991) is a standard reference on linear algebra. A statement of Weyl's theorem and its corollaries can be found in section 4.3 of the first volume (Horn and Johnson, 1985). The monograph by Bhatia (1997) is more advanced in nature, taking a functional-analytic perspective, and includes discussion of Lidskii's theorem (see section III.4). The notes by Carlen (2009) contain further background on trace inequalities, such as inequality (6.25).

Some classical papers on asymptotic random matrix theory include those by Wigner (1955; 1958), Marčenko and Pastur (1967), Pastur (1972), Wachter (1978) and Geman (1980). Mehta (1991) provides an overview of asymptotic random matrix theory, primarily from the physicist's perspective, whereas the book by Bai and Silverstein (2010) takes a more statistical perspective. The lecture notes of Vershynin (2011) focus on the non-asymptotic aspects of random matrix theory, as partially covered here. Davidson and Szarek (2001) describe the use of Sudakov–Fernique (Slepian) and Gordon inequalities in bounding expectations of random matrices; see also the earlier papers by Gordon (1985; 1986; 1987) and Szarek (1991). The results in Davidson and Szarek (2001) are for the special case of the standard Gaussian ensemble ($\mathbf{\Sigma} = \mathbf{I}_d$), but the underlying arguments are easily extended to the general case, as given here.

The proof of Theorem 6.5 is based on the lecture notes of Vershynin (2011). The underlying discretization argument is classical, used extensively in early work on random constructions in Banach space geometry (e.g., see the book by Pisier (1989) and references therein). Note that this discretization argument is the one-step version of the more sophisticated chaining methods described in Chapter 5.

Bounds on the expected operator norm of a random matrix follow a class of results known

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press

as non-commutative Bernstein inequalities, as derived initially by Rudelson (1999). Ahlswede and Winter (2002) developed techniques for matrix tail bounds based on controlling the matrix moment generating function, and exploiting the Golden–Thompson inequality. Other authors, among them Oliveira (2010), Gross (2011) and Recht (2011), developed various extensions and refinements of the original Ahlswede–Winter approach. Tropp (2010) introduced the idea of controlling the matrix generating function directly, and developed the argument that underlies Lemma 6.13. Controlling the moment generating function in this way leads to tail bounds involving the variance parameter $\sigma^2 := \frac{1}{n}|||\sum_{i=1}^n \mathrm{var}(\mathbf{Q}_i)|||_2$ as opposed to the potentially larger quantity $\tilde{\sigma}^2 := \frac{1}{n}\sum_{i=1}^n |||\mathrm{var}(\mathbf{Q}_i)|||_2$ that follows from the original Ahlswede–Winter argument. By the triangle inequality for the operator norm, we have $\sigma^2 \leq \tilde{\sigma}^2$, and the latter quantity can be substantially larger. Independent work by Oliveira (2010) also derived bounds involving the variance parameter σ^2, using a related technique that sharpened the original Ahlswede–Winter approach. Tropp (2010) also provides various extensions of the basic Bernstein bound, among them results for matrix martingales as opposed to the independent random matrices considered here. Mackey et al. (2014) show how to derive matrix concentration bounds with sharp constants using the method of exchangeable pairs introduced by Chatterjee (2007). Matrix tail bounds with refined forms of dimension dependence have been developed by various authors (Minsker, 2011; Hsu et al., 2012a); the specific sharpening sketched out in Example 6.19 and Exercise 6.12 is due to Minsker (2011).

For covariance estimation, Adamczak et al. (2010) provide sharp results on the deviation $|||\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}|||_2$ for distributions with sub-exponential tails. These results remove the superfluous logarithmic factor that arises from an application of Corollary 6.20 to a sub-exponential ensemble. Srivastava and Vershynin (2013) give related results under very weak moment conditions. For thresholded sample covariances, the first high-dimensional analyses were undertaken in independent work by Bickel and Levina (2008a) and El Karoui (2008). Bickel and Levina studied the problem under sub-Gaussian tail conditions, and introduced the row-wise sparsity model, defined in terms of the maximum ℓ_q-"norm" taken over the rows. By contrast, El Karoui imposed a milder set of moment conditions, and measured sparsity in terms of the growth rates of path lengths in the graph; this approach is essentially equivalent to controlling the operator norm $|||\mathbf{A}|||_2$ of the adjacency matrix, as in Theorem 6.23. The star graph is an interesting example that illustrates the difference between the row-wise sparsity model, and the operator norm approach.

An alternative model for covariance matrices is a banded decay model, in which entries decay according to their distance from the diagonal. Bickel and Levina (2008b) introduced this model in the covariance setting, and proposed a certain kind of tapering estimator. Cai et al. (2010) analyzed the minimax-optimal rates associated with this class of covariance matrices, and provided a modified estimator that achieves these optimal rates.

6.8 Exercises

Exercise 6.1 (Bounds on eigenvalues) Given two symmetric matrices **A** and **B**, show directly, without citing any other theorems, that

$$|\gamma_{\max}(\mathbf{A}) - \gamma_{\max}(\mathbf{B})| \leq |||\mathbf{A} - \mathbf{B}|||_2 \quad \text{and} \quad |\gamma_{\min}(\mathbf{A}) - \gamma_{\min}(\mathbf{B})| \leq |||\mathbf{A} - \mathbf{B}|||_2.$$

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press

Exercise 6.2 (Relations between matrix operator norms) For a rectangular matrix $\mathbf{A}$ with real entries and a scalar $q \in [1, \infty]$, the $(\ell_q \to \ell_q)$-operator norms are given by

$$|||\mathbf{A}|||_q = \sup_{\|x\|_q = 1} \|\mathbf{A}x\|_q.$$

(a) Derive explicit expressions for the operator norms $|||\mathbf{A}|||_2$, $|||\mathbf{A}|||_1$ and $|||\mathbf{A}|||_\infty$ in terms of elements and/or singular values of $\mathbf{A}$.

(b) Prove that $|||\mathbf{AB}|||_q \leq |||\mathbf{A}|||_q |||\mathbf{B}|||_q$ for any size-compatible matrices $\mathbf{A}$ and $\mathbf{B}$.

(c) For a square matrix $\mathbf{A}$, prove that $|||\mathbf{A}|||_2^2 \leq |||\mathbf{A}|||_1 |||\mathbf{A}|||_\infty$. What happens when $\mathbf{A}$ is symmetric?

Exercise 6.3 (Non-negative matrices and operator norms) Given two d-dimensional symmetric matrices $\mathbf{A}$ and $\mathbf{B}$, suppose that $0 \leq \mathbf{A} \leq \mathbf{B}$ in an elementwise sense (i.e., $0 \leq A_{j\ell} \leq B_{j\ell}$ for all $j, \ell = 1, \ldots, d$.)

(a) Show that $0 \leq \mathbf{A}^m \leq \mathbf{B}^m$ for all integers $m = 1, 2, \ldots$.

(b) Use part (a) to show that $|||\mathbf{A}|||_2 \leq |||\mathbf{B}|||_2$.

(c) Use a similar argument to show that $|||\mathbf{C}|||_2 \leq |||\, |\mathbf{C}| \,|||_2$ for any symmetric matrix $\mathbf{C}$, where $|\mathbf{C}|$ denotes the absolute value function applied elementwise.

Exercise 6.4 (Inequality for matrix exponential) Let $\mathbf{A} \in \mathcal{S}^{d \times d}$ be any symmetric matrix. Show that $\mathbf{I}_d + \mathbf{A} \preceq e^{\mathbf{A}}$. (*Hint:* First prove the statement for a diagonal matrix $\mathbf{A}$, and then show how to reduce to the diagonal case.)

Exercise 6.5 (Matrix monotone functions) A function $f \colon \mathcal{S}_+^{d \times d} \to \mathcal{S}_+^{d \times d}$ on the space of symmetric positive semidefinite matrices is said to be *matrix monotone* if

$$f(\mathbf{A}) \preceq f(\mathbf{B}) \qquad \text{whenever } \mathbf{A} \preceq \mathbf{B}.$$

Here $\preceq$ denotes the positive semidefinite ordering on $\mathcal{S}_+^{d \times d}$.

(a) Show by counterexample that the function $f(\mathbf{A}) = \mathbf{A}^2$ is *not* matrix monotone. (*Hint:* Note that $(\mathbf{A}+t\mathbf{C})^2 = \mathbf{A}^2 + t^2\mathbf{C}^2 + t(\mathbf{AC}+\mathbf{CA})$, and search for a pair of positive semidefinite matrices such that $\mathbf{AC} + \mathbf{CA}$ has a negative eigenvalue.)

(b) Show by counterexample that the matrix exponential function $f(\mathbf{A}) = e^{\mathbf{A}}$ is *not* matrix monotone. (*Hint:* Part (a) could be useful.)

(c) Show that the matrix logarithm function $f(\mathbf{A}) = \log \mathbf{A}$ is matrix monotone on the cone of strictly positive definite matrices. (*Hint:* You may use the fact that $g(\mathbf{A}) = \mathbf{A}^p$ is matrix monotone for all $p \in [0, 1]$.)

Exercise 6.6 (Variance and positive semidefiniteness) Recall that the variance of a symmetric random matrix $\mathbf{Q}$ is given by $\operatorname{var}(\mathbf{Q}) = \mathbb{E}[\mathbf{Q}^2] - (\mathbb{E}[\mathbf{Q}])^2$. Show that $\operatorname{var}(\mathbf{Q}) \succeq 0$.

Exercise 6.7 (Sub-Gaussian random matrices) Consider the random matrix $\mathbf{Q} = g\mathbf{B}$, where $g \in \mathbb{R}$ is a zero-mean σ-sub-Gaussian variable.

(a) Assume that g has a distribution symmetric around zero, and $\mathbf{B} \in \mathcal{S}^{d \times d}$ is a deterministic matrix. Show that $\mathbf{Q}$ is sub-Gaussian with matrix parameter $\mathbf{V} = c^2\sigma^2\mathbf{B}^2$, for some universal constant c.

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press

(b) Now assume that $\mathbf{B} \in \mathcal{S}^{d\times d}$ is random and independent of g, with $|||\mathbf{B}|||_2 \leq b$ almost surely. Prove that $\mathbf{Q}$ is sub-Gaussian with matrix parameter given by $\mathbf{V} = c^2\sigma^2 b^2 \mathbf{I}_d$.

Exercise 6.8 (Sub-Gaussian matrices and mean bounds) Consider a sequence of independent, zero-mean random matrices $\{\mathbf{Q}_i\}_{i=1}^n$ in $\mathcal{S}^{d\times d}$, each sub-Gaussian with matrix parameter $\mathbf{V}_i$. In this exercise, we provide bounds on the expected value of eigenvalues and operator norm of $\mathbf{S}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{Q}_i$.

(a) Show that $\mathbb{E}[\gamma_{\max}(\mathbf{S}_n)] \leq \sqrt{\frac{2\sigma^2 \log d}{n}}$, where $\sigma^2 = |||\frac{1}{n}\sum_{i=1}^n \mathbf{V}_i|||_2$.
(*Hint:* Start by showing that $\mathbb{E}[e^{\lambda\gamma_{\max}(\mathbf{S}_n)}] \leq de^{\frac{\lambda^2\sigma^2}{2n}}$.)

(b) Show that

$$\mathbb{E}\left[\left|\left|\left|\frac{1}{n}\sum_{i=1}^n \mathbf{Q}_i\right|\right|\right|_2\right] \leq \sqrt{\frac{2\sigma^2\log(2d)}{n}}. \tag{6.68}$$

Exercise 6.9 (Bounded matrices and Bernstein condition) Let $\mathbf{Q} \in \mathcal{S}^{d\times d}$ be an arbitrary symmetric matrix.

(a) Show that the bound $|||\mathbf{Q}|||_2 \leq b$ implies that $\mathbf{Q}^{j-2} \preceq b^{j-2}\mathbf{I}_d$.

(b) Show that the positive semidefinite order is preserved under left–right multiplication, meaning that if $\mathbf{A} \preceq \mathbf{B}$, then we also have $\mathbf{QAQ} \preceq \mathbf{QBQ}$ for any matrix $\mathbf{Q} \in \mathcal{S}^{d\times d}$.

(c) Use parts (a) and (b) to prove the inequality (6.30).

Exercise 6.10 (Tail bounds for non-symmetric matrices) In this exercise, we prove that a version of the tail bound (6.42) holds for general independent zero-mean matrices $\{\mathbf{A}_i\}_{i=1}^n$ that are almost surely bounded as $|||\mathbf{A}_i|||_2 \leq b$, as long as we adopt the new definition (6.43) of σ^2.

(a) Given a general matrix $\mathbf{A}_i \in \mathbb{R}^{d_1\times d_2}$, define a symmetric matrix of dimension $(d_1 + d_2)$ via

$$\mathbf{Q}_i := \begin{bmatrix} \mathbf{0}_{d_1\times d_2} & \mathbf{A}_i \\ \mathbf{A}_i^T & \mathbf{0}_{d_2\times d_1} \end{bmatrix}.$$

Prove that $|||\mathbf{Q}_i|||_2 = |||\mathbf{A}_i|||_2$.

(b) Prove that $|||\frac{1}{n}\sum_{i=1}^n \operatorname{var}(\mathbf{Q}_i)|||_2 \leq \sigma^2$ where σ^2 is defined in equation (6.43).

(c) Conclude that

$$\mathbb{P}\left[\left|\left|\left|\sum_{i=1}^n \mathbf{A}_i\right|\right|\right|_2 \geq n\delta\right] \leq 2(d_1 + d_2)e^{-\frac{n\delta^2}{2(\sigma^2+b\delta)}}. \tag{6.69}$$

Exercise 6.11 (Unbounded matrices and Bernstein bounds) Consider an independent sequence of random matrices $\{\mathbf{A}_i\}_{i=1}^n$ in $\mathbb{R}^{d_1\times d_2}$, each of the form $\mathbf{A}_i = g_i\mathbf{B}_i$, where $g_i \in \mathbb{R}$ is a zero-mean scalar random variable, and $\mathbf{B}_i$ is an independent random matrix. Suppose that $\mathbb{E}[g_i^j] \leq \frac{j!}{2} b_1^{j-2}\sigma^2$ for $j = 2, 3, \ldots$, and that $|||\mathbf{B}_i|||_2 \leq b_2$ almost surely.

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press

(a) For any $\delta > 0$, show that

$$\mathbb{P}\left[\left|\left|\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{A}_i\right|\right|\right|_2 \geq \delta\right] \leq (d_1 + d_2)e^{-\frac{n\delta^2}{2(\sigma^2 b_2^2 + b_1 b_2 \delta)}}.$$

(*Hint:* The result of Exercise 6.10(a) could be useful.)

(b) Show that

$$\mathbb{E}\left[\left|\left|\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{A}_i\right|\right|\right|_2\right] \leq \frac{2\sigma b_2}{\sqrt{n}}\left\{\sqrt{\log(d_1 + d_2)} + \sqrt{\pi}\right\} + \frac{4b_1 b_2}{n}\{\log(d_1 + d_2) + 1\}.$$

(*Hint:* The result of Exercise 2.8 could be useful.)

Exercise 6.12 (Sharpened matrix Bernstein inequality) In this exercise, we work through various steps of the calculation sketched in Example 6.19.

(a) Prove the bound (6.46).

(b) Show that for any symmetric zero-mean random matrix $\mathbf{Q}$ such that $|||\mathbf{Q}|||_2 \leq 1$ almost surely, the moment generating function is bounded as

$$\log \Psi_{\mathbf{Q}}(\lambda) \leq \underbrace{(e^{\lambda} - \lambda - 1)}_{\phi(\lambda)} \operatorname{var}(\mathbf{Q}).$$

(c) Prove the upper bound (6.47b).

Exercise 6.13 (Bernstein's inequality for vectors) In this exercise, we consider the problem of obtaining a Bernstein-type bound on random variable $\|\sum_{i=1}^{n} x_i\|_2$, where $\{x_i\}_{i=1}^{n}$ is an i.i.d. sequence of zero-mean random vectors such that $\|x_i\|_2 \leq 1$ almost surely, and $\operatorname{cov}(x_i) = \mathbf{\Sigma}$. In order to do so, we consider applying either Theorem 6.17 or the bound (6.48) to the $(d+1)$-dimensional symmetric matrices

$$\mathbf{Q}_i := \begin{bmatrix} 0 & x_i^{\mathrm{T}} \\ x_i & \mathbf{0}_d \end{bmatrix}.$$

Define the matrix $\mathbf{V}_n = \sum_{i=1}^{n} \operatorname{var}(\mathbf{Q}_i)$.

(a) Show that the best bound obtainable from Theorem 6.17 will have a pre-factor of the form $\operatorname{rank}(\mathbf{\Sigma}) + 1$, which can be as large as $d + 1$.

(b) By way of contrast, show that the bound (6.48) yields a dimension-independent pre-factor of 2.

Exercise 6.14 (Random packings) The goal of this exercise is to prove that there exists a collection of vectors $\mathcal{P} = \{\theta^1, \ldots, \theta^M\}$ belonging to the sphere $\mathbb{S}^{d-1}$ such that:

(a) the set $\mathcal{P}$ forms a $1/2$-packing in the Euclidean norm;

(b) the set $\mathcal{P}$ has cardinality $M \geq e^{c_0 d}$ for some universal constant c_0;

(c) the inequality $|||\frac{1}{M}\sum_{j=1}^{M}(\theta^j \otimes \theta^j)|||_2 \leq \frac{2}{d}$ holds.

(*Note:* You may assume that d is larger than some universal constant so as to avoid annoying subcases.)

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press

Exercise 6.15 (Estimation of diagonal covariances) Let $\{x_i\}_{i=1}^n$ be an i.i.d. sequence of d-dimensional vectors, drawn from a zero-mean distribution with diagonal covariance matrix $\mathbf{\Sigma} = \mathbf{D}$. Consider the estimate $\widehat{\mathbf{D}} = \mathrm{diag}(\widehat{\mathbf{\Sigma}})$, where $\widehat{\mathbf{\Sigma}}$ is the usual sample covariance matrix.

(a) When each vector x_i is sub-Gaussian with parameter at most σ, show that there are universal positive constants c_j such that

$$\mathbb{P}\left[|||\widehat{\mathbf{D}} - \mathbf{D}|||_2/\sigma^2 \geq c_0\sqrt{\frac{\log d}{n}} + \delta\right] \leq c_1 e^{-c_2 n \min\{\delta, \delta^2\}}, \qquad \text{for all } \delta > 0.$$

(b) Instead of a sub-Gaussian tail condition, suppose that for some even integer $m \geq 2$, there is a universal constant K_m such that

$$\underbrace{\mathbb{E}[(x_{ij}^2 - \Sigma_{jj})^m]}_{\|x_{ij}^2 - \Sigma_{jj}\|_m^m} \leq K_m \qquad \text{for each } i = 1, \ldots, n \text{ and } j = 1, \ldots, d.$$

Show that

$$\mathbb{P}\left[|||\widehat{\mathbf{D}} - \mathbf{D}|||_2 \geq 4\delta\sqrt{\frac{d^{2/m}}{n}}\right] \leq K'_m\left(\frac{1}{2\delta}\right)^m \qquad \text{for all } \delta > 0,$$

where K'_m is another universal constant.

Hint: You may find Rosenthal's inequality useful: given zero-mean independent random variables Z_i such that $\|Z_i\|_m < +\infty$, there is a universal constant C_m such that

$$\left\|\sum_{i=1}^n Z_i\right\|_m \leq C_m\left\{\left(\sum_{i=1}^n \mathbb{E}[Z_i^2]\right)^{1/2} + \left(\sum_{i=1}^n \mathbb{E}[|Z_i|^m]\right)^{1/m}\right\}.$$

Exercise 6.16 (Graphs and adjacency matrices) Let G be a graph with maximum degree $s - 1$ that contains an s-clique. Letting $\mathbf{A}$ denote its adjacency matrix (defined with ones on the diagonal), show that $|||\mathbf{A}|||_2 = s$.

https://doi.org/10.1017/9781108627771.006 Published online by Cambridge University Press