# Chapter 2
# Sums of Matrix-Valued Random Variables

This chapter gives an exhaustive treatment of the line of research for sums of matrix-valued random matrices. We will present eight different derivation methods in this context of matrix Laplace transform method. The emphasis is placed on the methods that will be hopefully useful to some engineering applications. Although powerful, the methods are elementary in nature. It is remarkable that some modern results on matrix completion can be simply derived, by using the framework of sums of matrix-valued random matrices. The treatment here is self-contained. All the necessary tools are developed in Chap. 1. The contents of this book are complementary to our book [5]. We have a small overlapping on the results of [36].

In this chapter, the classical, commutative theory of probability is generalized to the more general theory of non-communicative probability. Non-communicative algebras of random variables ("observations") and their expectations (or "trace") are built. *Matrices or operators takes the role of scalar random variables and the trace takes the role of expectation*. This is very similar to free probability [9].

## 2.1 Methodology for Sums of Random Matrices

The theory of real random variables provides the framework of much of modern probability theory [8], such as laws of large numbers, limit theorems, and probability estimates for "deviations", when sums of independent random variables are involved. However, some authors have started to develop analogous theories for the case that the algebraic structure of the reals is substituted by more general structures such as groups, vector spaces, etc., see for example [90].

In a remarkable work [36], Ahlswede and Winter has laid the ground for the fundamentals of a theory of (self-adjoint) operator valued random variables. There, the large deviation bounds are derived. A self-adjoint operator includes finite dimensions (often called Hermitian matrix) and infinite dimensions. For the purpose of this book, finite dimensions are sufficient. We will prefer Hermitian matrix.

R. Qiu and M. Wicks, *Cognitive Networked Sensing and Big Data*,
DOI 10.1007/978-1-4614-4544-9_2,

To extend the theory from scalars to the matrices, the fundamental difficulty arises from that fact, in general, two matrices are not commutative. For example, $\mathbf{AB} \neq \mathbf{BA}$. The functions of a matrix can be defined; for example, the matrix exponential is defined [20] as $e^{\mathbf{A}}$. As expected, $e^{\mathbf{AB}} \neq e^{\mathbf{BA}}$, although a scalar exponential has the elementary property $e^{ab} = e^{ba}$, for two scalars $a, b$. Fortunately, we have the Golden-Thompson inequality that has the limited replacement for the above elementary property of the scalar exponential. The Golden-Thompson inequality

$$\mathrm{Tr}\left(e^{\mathbf{A}+\mathbf{B}}\right) \leq \mathrm{Tr}\left(e^{\mathbf{A}}e^{\mathbf{B}}\right),$$

for Hermitian matrices $\mathbf{A}, \mathbf{B}$, is the most complicate result that we will use.

Through the spectral mapping theorem, the eigenvalues of arbitrary matrix function $f(\mathbf{A})$, are $f(\lambda_i)$ where $\lambda_i$ is the $i$-th eigenvalue of $\mathbf{A}$. In particular, for $f(x) = e^x$ for a scalar $x$; the eigenvalues of $e^{\mathbf{A}}$ are $e^{\lambda_i}$, which is, of course, positive (i.e., $e^{\lambda_i} > 0$). In other words, the matrix exponential $e^{\mathbf{A}}$ is ALWAYS positive semidefinite for an arbitrary matrix $\mathbf{A}$. The positive real numbers have a lot of special structures to exploit, compared with arbitrary real numbers. The elementary fact motivates the wide use of positive semidefinite (PSD) matrices, for example, convex optimization and quantum information theory. Through the spectral mapping theorem, all the eigenvalues of positive semidefinite matrices are nonnegative.

For a sequence of scalar random variables (real or complex numbers), $x_1, \ldots, x_n$, we can study its convergence by studying the so-called partial sum $S_n = x_1 + \ldots + x_n = \sum_{i=1}^{n} x_i$. We say the sequence converges to a limit value $S = \mathbb{E}[x]$, if there exists a limit $S$ as $n \to \infty$. In analogy with the scalar counterparts, we can similarly define

$$\mathbf{S}_n = \mathbf{X}_1 + \ldots + \mathbf{X}_n = \sum_{i=1}^{n} \mathbf{X}_i,$$

for a sequence of Hermitian matrices, $\mathbf{X}_1, \ldots, \mathbf{X}_n$. We say the sequence converges to a limit matrix $\mathbf{S} = \mathbb{E}[\mathbf{X}]$, if there exists a limit $\mathbf{S}$ as $n \to \infty$.

One nice thing about the positive number is the ordering. When $a = 0.4$ and $b = 0.5$, we can say $a < b$. In analogy, we say the partial order $\mathbf{A} \leq \mathbf{B}$ if all the eigenvalues of $\mathbf{B} - \mathbf{A}$ are nonnegative, which is equivalent to say $\mathbf{B} - \mathbf{A}$ is positive semidefinite matrix. Since a matrix exponential is $e^{\mathbf{A}}$ is always positive semidefinite for an arbitrary matrix $\mathbf{A}$, we can instead study $e^{\mathbf{A}} \leq e^{\mathbf{B}}$, to infer about the partial order $\mathbf{A} \leq \mathbf{B}$. The function $x \to e^{sx}$ is monotone, non-decreasing and positive for all $s \geq 0$. we can, by the spectral mapping theorem to study their eigenvalues which are scalar random variables. Thus a matrix-valued random variable is converted into a scalar-valued random variable, by using the bridge of the spectral mapping theorem. For our interest, what matters is the spectrum (spec$(\mathbf{A})$, the set of all eigenvalues).

In summary, the sums of random matrices are of elementary nature. We emphasize the fundamental contribution of Ahlswede and Winter [36] since their work has triggered a snow ball of this line of research.

## 2.2 Matrix Laplace Transform Method

Due to the basic nature of sums of random matrices, we give several versions of the theorems and their derivations. Although essentially their techniques are equivalent, the assumptions and arguments are sufficiently different to justify the space. The techniques for handling matrix-valued random variables are very subtle; it is our intention to give an exhaustive survey of these techniques. Even a seemingly small twist of the problems can cause a lot of technical difficulties. These presentations serve as examples to illustrate the key steps. Repetition is the best teacher—practice makes it perfect. This is the rationale behind this chapter. It is hoped that the audience pays attention to the methods, not the particular derived inequalities.

The Laplace transform method is the standard technique for the scalar-valued random variables; it is remarkable that this method can be extended to the matrix setting. We argue that this is a break-through in studying the matrices concentration. This method is used as a thread to tie together all the surveyed literature. For completion, we run the risk of "borrowing" too much from the cited references. Here we give credit to those cited authors. We try our best to add more details about their arguments with the hope of being more accessible.

### *2.2.1 Method 1—Harvey's Derivation*

The presentation here is essentially the same as [91,92] whose style is very friendly and accessible. We present Harvey's version first.

#### 2.2.1.1 The Ahlswede-Winter Inequality

Let $\mathbf{X}$ be a random $d \times d$ matrix, i.e., a matrix whose entries are all random variables. We define $\mathbb{E}\mathbf{X}$ to be the matrix whose entries are the expectation of the entries of $\mathbf{X}$. Since expectation and trace are both linear, they commute:

$$\mathbb{E}\left[\operatorname{Tr}\mathbf{X}\right] \triangleq \sum_{\mathbf{A}} \mathbb{P}\left(\mathbf{X} = \mathbf{A}\right) \cdot \sum_i A_{i,i} = \sum_i \sum_{\mathbf{A}} \mathbb{P}\left(\mathbf{X} = \mathbf{A}\right) \cdot A_{i,i}$$

$$= \sum_i \sum_a \mathbb{P}\left(X_{i,i} = a\right) \cdot a = \sum_i \mathbb{E}\left(X_{i,i}\right) = \operatorname{Tr}\left(\mathbb{E}\mathbf{X}\right).$$

Let $\mathbf{X}_1, \cdots, \mathbf{X}_n$ be random, *symmetric*[1] matrices of size $d \times d$. Define the partial sums

$$\mathbf{S}_n = \mathbf{X}_1 + \cdots + \mathbf{X}_n = \sum_{i=1}^{n} \mathbf{X}_i.$$

$\mathbf{A} \geq 0$ is equivalent to saying that all eigenvalues of $\mathbf{A}$ are nonnegative, i.e., $\lambda_i(\mathbf{A}) \geqslant 0$. We would like to analyze the probability that eigenvalues of $\mathbf{S}_n$ are at most $t$, i.e., $\mathbf{S}_n \leq t\mathbf{I}$. This is equivalent to the event that all eigenvalues of $e^{\mathbf{S}_n}$ are at most $e^{\lambda t}$, i.e., $e^{\mathbf{S}_n \lambda} \leqslant e^{\lambda t \mathbf{I}}$. If this event fails to hold, then certainly $\operatorname{Tr} e^{\mathbf{S}_n \lambda} > \operatorname{Tr} e^{\lambda t \mathbf{I}}$, since all eigenvalues of $e^{\mathbf{S}_n}$ are non-negative. Thus, we have argued that

$$\begin{aligned}
&\Pr\left[\text{some eigenvalues of matrix } \mathbf{S}_n \text{ is greater than t}\right] \\
&\leqslant \mathbb{P}\left(\operatorname{Tr} e^{\mathbf{S}_n t} > \operatorname{Tr} e^{\lambda t \mathbf{I}}\right) \\
&\leqslant \mathbb{E}\left(\operatorname{Tr} e^{\mathbf{S}_n t}\right) / e^{\lambda t},
\end{aligned} \tag{2.1}$$

by Markov's inequality. Now, as in the proof of the Chernoff bound, we want to bound this expectation by a product of expectations, which will lead to an exponentially decreasing tail bound. This is where the Golden-Thompson inequality is needed.

$$\begin{aligned}
&\mathbb{E}\left(\operatorname{Tr} e^{\mathbf{S}_n \lambda}\right) = \mathbb{E}\left(\operatorname{Tr} e^{\lambda \mathbf{X}_n + \lambda \mathbf{S}_{n-1}}\right) \text{(since } \mathbf{S}_n = \mathbf{X}_n + \mathbf{S}_{n-1}\text{)} \\
&\leqslant \mathbb{E}\left[\operatorname{Tr}\left(e^{\lambda \mathbf{X}_n} \cdot e^{\lambda \mathbf{S}_{n-1}}\right)\right] \text{(by Golden - Thompson inequality)} \\
&= \mathbb{E}_{\mathbf{X}_1, \cdots, \mathbf{X}_{n-1}}\left\{\mathbb{E}_{\mathbf{X}_n}\left[\operatorname{Tr}\left(e^{\lambda \mathbf{X}_n} \cdot e^{\lambda \mathbf{S}_{n-1}}\right)\right]\right\} \text{(since the } \mathbf{X}_i\text{'s are mutually independent)} \\
&= \mathbb{E}_{\mathbf{X}_1, \cdots, \mathbf{X}_{n-1}}\left\{\operatorname{Tr}\left[\mathbb{E}_{\mathbf{X}_n}\left(e^{\lambda \mathbf{X}_n} \cdot e^{\lambda \mathbf{S}_{n-1}}\right)\right]\right\} \text{(since trace and expectation commute)} \\
&= \mathbb{E}_{\mathbf{X}_1, \cdots, \mathbf{X}_{n-1}}\left\{\operatorname{Tr}\left[\mathbb{E}_{\mathbf{X}_n}\left(e^{\lambda \mathbf{X}_n}\right) \cdot e^{\lambda \mathbf{S}_{n-1}}\right]\right\} \text{(since } \mathbf{X}_n \text{ and } \mathbf{S}_{n-1} \text{are independent)} \\
&= \mathbb{E}_{\mathbf{X}_1, \cdots, \mathbf{X}_{n-1}}\left[\left\|\mathbb{E}_{\mathbf{X}_n}\left(e^{\lambda \mathbf{X}_n}\right)\right\| \cdot \operatorname{Tr}\left(e^{\lambda \mathbf{S}_{n-1}}\right)\right] \text{(by Corollary of trace-norm property)} \\
&= \left\|\mathbb{E}_{\mathbf{X}_n}\left(e^{\lambda \mathbf{X}_n}\right)\right\| \cdot \mathbb{E}_{\mathbf{X}_1, \cdots, \mathbf{X}_{n-1}}\left[\operatorname{Tr}\left(e^{\lambda \mathbf{S}_{n-1}}\right)\right]
\end{aligned} \tag{2.2}$$

Applying this inequality inductively, we get

$$\mathbb{E}\left(\operatorname{Tr} e^{\mathbf{S}_n \lambda}\right) \leqslant \prod_{i=1}^{n} \left\|\mathbb{E}_{\mathbf{X}_i}\left(e^{\lambda \mathbf{X}_i}\right)\right\| \cdot \operatorname{Tr}\left(e^{\lambda \mathbf{0}}\right) = \prod_{i=1}^{n} \left\|\mathbb{E}\left(e^{\lambda \mathbf{X}_i}\right)\right\| \cdot \operatorname{Tr}\left(e^{\lambda \mathbf{0}}\right),$$

[1]The assumption symmetric matrix is too strong for many applications. Since we often deal with complex entries, the assumption of Hermitian matrix is reasonable. This is the fatal flaw of this version. Otherwise, it is very useful.

where $\mathbf{0}$ is the zero matrix of size $d \times d$. So $e^{\lambda \mathbf{0}} = \mathbf{I}$ and $\operatorname{Tr}(\mathbf{I}) = d$, where $\mathbf{I}$ is the identity matrix whose diagonal are all 1. Therefore,

$$\mathbb{E}\left(\operatorname{Tr} e^{\mathbf{S}_n \lambda}\right) \leqslant d \cdot \prod_{i=1}^{n} \left\| \mathbb{E}\left(e^{\lambda \mathbf{X}_i}\right) \right\|.$$

Combining this with (2.1), we obtain

$$\Pr\left[\text{some eigenvalues of matrix } \mathbf{S}_n \text{ is greater than t}\right] \leqslant d e^{-\lambda t} \prod_{i=1}^{n} \left\| \mathbb{E}\left(e^{\lambda \mathbf{X}_i}\right) \right\|.$$

We can also bound the probability that any eigenvalue of $\mathbf{S}_n$ is less than $-t$ by applying the same argument to $-\mathbf{S}_n$. This shows that the probability that any eigenvalue of $\mathbf{S}_n$ lies outside $[-t, t]$ is

$$\mathbb{P}\left(\|\mathbf{S}_n\| > t\right) \leqslant d e^{-\lambda t} \left\{ \prod_{i=1}^{n} \left\| \mathbb{E}\left(e^{\lambda \mathbf{X}_i}\right) \right\| + \prod_{i=1}^{n} \left\| \mathbb{E}\left(e^{-\lambda \mathbf{X}_i}\right) \right\| \right\}. \tag{2.3}$$

This is the basis inequality. Much like the Chernoff bound, numerous variations and generalizations are possible. Two useful versions are stated here without proof.

**Theorem 2.2.1.** *Let $\mathbf{Y}$ be a random, symmetric, positive semi-definite $d \times d$ matrix such that $\mathbb{E}[\mathbf{Y}] = \mathbf{I}$. Suppose $\|Y\| \leq R$ for some fixed scalar $R \geq 1$. Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_k$ be independent copies of $\mathbf{Y}$ (i.e., independently sampled matrices with the same distribution as $\mathbf{Y}$). For any $\varepsilon \in (0, 1)$, we have*

$$\mathbb{P}\left[ (1-\varepsilon)\,\mathbf{I} \leqslant \frac{1}{k} \sum_{i=1}^{k} \mathbf{Y}_i \leqslant (1+\varepsilon)\,\mathbf{I} \right] \geqslant 1 - 2d \cdot \exp\left(-\varepsilon^2 k / 4R\right).$$

*This event is equivalent to the sample average $\frac{1}{k} \sum\limits_{i=1}^{k} \mathbf{Y}_i$ having minimum eigenvalue at least $1-\varepsilon$ and maximum eigenvalue at most $1+\varepsilon$.*

*Proof.* See [92]. □

**Corollary 2.2.2.** *Let $\mathbf{Z}$ be a random, symmetric, positive semi-definite $d \times d$ matrix. Define $\mathbf{U} = \mathbb{E}[\mathbf{Z}]$ and suppose $\mathbf{Z} \leq R \cdot \mathbf{U}$ for some scalar $R \geq 1$. Let $\mathbf{Z}_1, \ldots, \mathbf{Z}_k$ be independent copies of $\mathbf{Z}$ (i.e., independently sampled matrices with the same distribution as $\mathbf{Z}$). For any $\varepsilon \in (0, 1)$, we have*

$$\mathbb{P}\left[ (1-\varepsilon)\,\mathbf{U} \leqslant \frac{1}{k} \sum_{i=1}^{k} \mathbf{Z}_i \leqslant (1+\varepsilon)\,\mathbf{U} \right] \geqslant 1 - 2d \cdot \exp\left(-\varepsilon^2 k / 4R\right).$$

*Proof.* See [92]. □

#### 2.2.1.2 Rudelson's Theorem

In this section, we use how the Ahlswede-Winter inequality is used to prove a concentration inequality for random vectors due to Rudelson. His original proof was quite different [93].

The motivation for Rudelson's inequality comes from the problem of approximately computing the volume of a convex body. When solving this problem, a convenient first step is to transform the body into the "isotropic position", which is a technical way of saying "roughly like the unit sphere." To perform this first step, one requires a concentration inequality for randomly sampled vectors, which is provided by Rudelson's theorem.

**Theorem 2.2.3 (Rudelson's Theorem [93]).** *Let* $\mathbf{x} \in \mathbb{R}^d$ *be a random vector such that* $\mathbb{E}\left(\mathbf{x}\mathbf{x}^T\right) = \mathbf{I}$. *Suppose* $\|\mathbf{x}\| \leqslant R$. *Let* $\mathbf{x}_1, \ldots, \mathbf{x}_n$ *be independent copies of* $\mathbf{x}$. *For any* $\varepsilon \in (0,1)$, *we have*

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T - \mathbf{I}\right\| > \varepsilon\right) \leqslant 2d \cdot \exp\left(-\varepsilon^2 n/4R^2\right).$$

Note that $R \geqslant \sqrt{d}$ because

$$d = \mathrm{Tr}\,\mathbf{I} = \mathrm{Tr}\left[\mathbb{E}\left(\mathbf{x}\mathbf{x}^T\right)\right] = \mathbb{E}\left[\mathrm{Tr}\left(\mathbf{x}\mathbf{x}^T\right)\right] = \mathbb{E}\left[\mathbf{x}^T\mathbf{x}\right],$$

since $\mathrm{Tr}\left(\mathbf{AB}\right) = \mathrm{Tr}\left(\mathbf{BA}\right)$.

*Proof.* We apply the Ahlswede-Winter inequality with the rank-1 matrix $\mathbf{X}_i$

$$\mathbf{X}_i = \frac{1}{2R^2}\left(\mathbf{x}_i\mathbf{x}_i^T - \mathbb{E}\left[\mathbf{x}_i\mathbf{x}_i^T\right]\right) = \frac{1}{2R^2}\left(\mathbf{x}_i\mathbf{x}_i^T - \mathbf{I}\right).$$

Note that $\mathbb{E}\mathbf{X}_i = \mathbf{0}$, $\|\mathbf{X}_i\| \leqslant 1$, and

$$\begin{aligned}
\mathbb{E}\left[\mathbf{X}_i^2\right] &= \frac{1}{4R^4}\mathbb{E}\left[\left(\mathbf{x}_i\mathbf{x}_i^T - \mathbf{I}\right)^2\right] \\
&= \frac{1}{4R^4}\left\{\mathbb{E}\left[\left(\mathbf{x}_i\mathbf{x}_i^T\right)^2\right] - \mathbf{I}\right\} \text{ (since } \mathbb{E}\left[\mathbf{x}_i\mathbf{x}_i^T\right] = \mathbf{I}\text{)} \\
&\leqslant \frac{1}{4R^4}\mathbb{E}\left[\left(\mathbf{x}_i^T\mathbf{x}_i\right)\left(\mathbf{x}_i\mathbf{x}_i^T\right)\right] \\
&\leqslant \frac{R^2}{4R^4}\mathbb{E}\left[\mathbf{x}_i\mathbf{x}_i^T\right] \text{ (since } \|\mathbf{x}_i\| \leqslant R\text{)} \\
&= \frac{\mathbf{I}}{4R^2}.
\end{aligned} \tag{2.4}$$

Now using Claim 1.4.5 together with the inequalities

$$1 + y \leqslant e^y, \forall y \in \mathbb{R}$$
$$e^y \leqslant 1 + y + y^2, \forall y \in [-1, 1].$$

Since $\|\mathbf{X}_i\| \leqslant 1$, for any $\lambda \in [0, 1]$, we have $e^{\lambda \mathbf{X}_i} \leqslant \mathbf{I} + \lambda \mathbf{X}_i + \lambda^2 \mathbf{X}_i^2$, and so

$$\mathbb{E}\left[e^{\lambda \mathbf{X}_i}\right] \leqslant \mathbb{E}\left[\mathbf{I} + \lambda \mathbf{X}_i + \lambda^2 \mathbf{X}_i^2\right] \leqslant \mathbf{I} + \lambda^2 \mathbb{E}\left[\mathbf{X}_i^2\right]$$
$$\leqslant e^{\lambda^2 \mathbb{E}\left[\mathbf{X}_i^2\right]} \leqslant e^{\lambda^2/4R^2 \mathbf{I}},$$

by Eq. (2.4). Thus, $\left\|\mathbb{E}\left[e^{\lambda \mathbf{X}_i}\right]\right\| \leqslant e^{\lambda^2/4R^2}$. The same analysis also shows that $\left\|\mathbb{E}\left[e^{-\lambda \mathbf{X}_i}\right]\right\| \leqslant e^{\lambda^2/4R^2}$. Substituting this into Eq. (2.3), we obtain

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n} \frac{1}{2R^2}\left(\mathbf{x}_i \mathbf{x}_i^T - \mathbf{I}\right)\right\| > t\right) \leqslant 2d \cdot e^{-\lambda t} \prod_{i=1}^{n} e^{\lambda^2/4R^2} = 2d \cdot \exp\left(-\lambda t + n\lambda^2/4R^2\right).$$

Substituting $t = n\varepsilon/2R^2$ and $\lambda = \varepsilon$ proves the theorem. □

### 2.2.2 Method 2—Vershynin's Derivation

We give the derivation method, taken from [35], by Vershynin.

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be independent random $d \times d$ real matrices, and let

$$\mathbf{S} = \mathbf{X}_1 + \cdots + \mathbf{X}_n.$$

We will be interested in the magnitude of the derivation $\|\mathbf{S}_n - \mathbb{E}\mathbf{S}_n\|$ in the operator norm $\|\cdot\|$.

Now we try to generalize the method of Sect. 1.4.10 when $\mathbf{X}_i \in \mathbb{M}_d$ are independent random matrices of mean zero, where $\mathbb{M}_d$ denotes the class of symmetric $d \times d$ matrices.

For, $\mathbf{A} \in \mathbb{M}_d$, the matrix exponential $e^{\mathbf{A}}$ is defined as usual by Taylor series. $e^{\mathbf{A}}$ has the same eigenvectors as $\mathbf{A}$, and eigenvalues $e^{\lambda_i(\mathbf{A})} > 0$. The partial order $\mathbf{A} \geq \mathbf{B}$ means $\mathbf{A} - \mathbf{B} \geq 0$, i.e., $\mathbf{A} - \mathbf{B}$ is positive semi-definite (their eigenvalues are non-negative). By using the exponential function of $\mathbf{A}$, we deal with the positive semi-definite matrix which has a fundamental structure to exploit.

The non-trivial part is that, in general,

$$e^{\mathbf{A}+\mathbf{B}} \neq e^{\mathbf{A}} e^{\mathbf{B}}.$$

However, the famous Golden-Thompson's inequality [94] sates that

$$\mathrm{Tr}\left(e^{\mathbf{A}+\mathbf{B}}\right) \leqslant \mathrm{Tr}\left(e^{\mathbf{A}} e^{\mathbf{B}}\right)$$

holds for arbitrary $\mathbf{A}, \mathbf{B} \in \mathbb{M}_d$ (and in fact for arbitrary unitary-invariant norm replacing the trace [23]). Therefore, for $\mathbf{S}_n = \mathbf{X}_1 + \cdots + \mathbf{X}_n = \sum_{i=1}^{n} \mathbf{X}_i$ and for $\mathbf{I}$ being the identify matrix on $\mathbb{M}_d$, we have

$$p \triangleq \mathbb{P}\left(\mathbf{S}_n \nleqslant t\mathbf{I}\right) = \mathbb{P}\left(e^{\lambda \mathbf{S}_n} \nleqslant e^{\lambda t \mathbf{I}}\right) \leqslant \mathbb{P}\left(\operatorname{Tr} e^{\lambda \mathbf{S}_n} > e^{\lambda t}\right) \leqslant e^{-\lambda t} \mathbb{E}\operatorname{Tr}\left(e^{\lambda \mathbf{S}_n}\right).$$

This estimate is not sharp: $e^{\lambda \mathbf{S}_n} \nleqslant e^{t\mathbf{I}}$ means the biggest eigenvalue of $e^{\lambda \mathbf{S}_n}$ exceeds $e^{\lambda t}$, while $\operatorname{Tr} e^{\lambda \mathbf{S}_n} > e^{\lambda t}$ means that the sum of all $d$ eigenvalues exceeds the same.

Since $\mathbf{S}_n = \mathbf{X}_n + \mathbf{S}_{n-1}$, we use the Golden-Thomas's inequality to separate the last term from the sum:

$$\mathbb{E}\operatorname{Tr}\left(e^{\lambda \mathbf{S}_n}\right) \leqslant \mathbb{E}\operatorname{Tr}\left(e^{\lambda \mathbf{X}_n} e^{\lambda \mathbf{S}_{n-1}}\right).$$

Now, using independence and that $\mathbb{E}$ and trace commute, we continue to write

$$= \mathbb{E}_{n-1}\operatorname{Tr}\left[\left(\mathbb{E}_n e^{\lambda \mathbf{X}_n}\right) \cdot e^{\lambda \mathbf{S}_{n-1}}\right] \leqslant \left\|\mathbb{E}_n e^{\lambda \mathbf{X}_n}\right\| \cdot \mathbb{E}_{n-1}\operatorname{Tr}\left(e^{\lambda \mathbf{S}_{n-1}}\right),$$

since

$$\operatorname{Tr}\left(\mathbf{AB}\right) \leqslant \|\mathbf{A}\| \operatorname{Tr}\left(\mathbf{B}\right),$$

for $\mathbf{A}, \mathbf{B} \in \mathbb{M}_d$.

Continuing by induction, we reach (since $\operatorname{Tr}\mathbf{I} = d$) to

$$\mathbb{E}\operatorname{Tr}\left(e^{\lambda \mathbf{S}_n}\right) \leqslant d \cdot \prod_{i=1}^{n} \mathbb{E} e^{\lambda \mathbf{X}_i}.$$

We have proved that

$$\mathbb{P}\left(\mathbf{S}_n \nleqslant t\mathbf{I}\right) \leqslant d \cdot \prod_{i=1}^{n} \mathbb{E} e^{\lambda \mathbf{X}_i}.$$

Repeating for $-\mathbf{S}_n$ and using that $-t\mathbf{I} \leqslant \mathbf{S}_n \leqslant t\mathbf{I}$ is equivalent to $\|\mathbf{S}_n\| \leqslant t$ we have shown that

$$\mathbb{P}\left(\|\mathbf{S}_n\| > t\right) \leqslant 2de^{-\lambda t} \cdot \prod_{i=1}^{n} \left\|\mathbb{E} e^{\lambda \mathbf{X}_i}\right\|. \tag{2.5}$$

As in the real valued case, full independence is never needed in the above argument. It works out well for martingales.

**Theorem 2.2.4 (Chernoff-type inequality).** *Let $\mathbf{X}_i \in \mathbb{M}_d$ be independent mean zero random matrices, $\|\mathbf{X}_i\| \leqslant 1$ for all $i$ almost surely. Let*

$$\mathbf{S}_n = \mathbf{X}_1 + \cdots + \mathbf{X}_n = \sum_{i=1}^{n} \mathbf{X}_i,$$

$$\sigma^2 = \sum_{i=1}^{n} \|\mathrm{var}\, \mathbf{X}_i\|.$$

*Then, for every* $t > 0$*, we have*

$$\mathbb{P}\left(\|\mathbf{S}_n\| > t\right) \leqslant d \cdot \max\left(e^{-t^2/4\sigma^2}, e^{-t/2}\right).$$

To prove this theorem, we have to estimate (2.5). The standard estimate

$$1 + y \leqslant e^y \leqslant 1 + y + y^2$$

is valid for real number $y \in [-1, 1]$ (actually a bit beyond) [95]. From the two bounds, we get (replacing $y$ with $\mathbf{Y}$)

$$\mathbf{I} + \mathbf{Y} \leqslant e^{\mathbf{Y}} \leqslant \mathbf{I} + \mathbf{Y} + \mathbf{Y}^2$$

Using the bounds twice (first the upper bound and then the lower bound), we have

$$\mathbb{E}e^{\mathbf{Y}} \leqslant \mathbb{E}\left(\mathbf{I} + \mathbf{Y} + \mathbf{Y}^2\right) = \mathbf{I} + \mathbb{E}\left(\mathbf{Y}^2\right) \leqslant e^{\mathbb{E}\left(\mathbf{Y}^2\right)}.$$

Let $0 < \lambda \leq 1$. Therefore, by the Theorem's hypothesis,

$$\left\|\mathbb{E}e^{\lambda \mathbf{X}_i}\right\| \leqslant \left\|e^{\lambda^2 \mathbb{E}\left(\mathbf{X}_i^2\right)}\right\| = e^{\lambda^2 \left\|\mathbb{E}\left(\mathbf{X}_i^2\right)\right\|}.$$

Hence by (2.5),

$$\mathbb{P}\left(\|\mathbf{S}_n\| > t\right) \leqslant 2d \cdot e^{-\lambda t + \lambda^2 \sigma^2}.$$

With the optimal choice of $\lambda = \min\left(t/2\sigma^2, 1\right)$, the conclusion of the Theorem follows.

Does the Theorem hold for $\sigma^2$ replaced by $\sum\limits_{i=1}^{n} \left\|\mathbb{E}\left(\mathbf{X}_i^2\right)\right\|$?

**Corollary 2.2.5.** *Let* $\mathbf{X}_i \in \mathbb{M}_d$ *be independent mean zero random matrices,* $\|\mathbf{X}_i\| \leqslant 1$ *for all* $i$ *almost surely. Let*

$$\mathbf{S}_n = \mathbf{X}_1 + \cdots + \mathbf{X}_n = \sum_{i=1}^{n} \mathbf{X}_i, \quad E = \sum_{i=1}^{n} \|\mathbb{E}\mathbf{X}_i\|.$$

*Then, for every* $\varepsilon \in (0,1)$*, we have*

$$\mathbb{P}\left(\|\mathbf{S}_n - \mathbb{E}\mathbf{S}_n\| > \varepsilon E\right) \leqslant d \cdot e^{-\varepsilon^2 E/4}.$$

### *2.2.3 Method 3—Oliveria's Derivation*

Consider the random matrix $\mathbf{Z}_n$. We closely follow Oliveria [38] whose exposition is highly accessible. In particular, he reviews all the needed theorems, all of those are collected in Chap. 1 for easy reference. In this subsection, the matrices are assumed to be $d \times d$ Hermitian matrices, that is, $\mathbf{A} \in \mathbb{C}_{\text{Herm}}^{d\times d}$, where $\mathbb{C}_{\text{Herm}}^{d\times d}$ is the set of $d \times d$ Hermitian matrices.

#### 2.2.3.1 Bernstein Trick

The usual Bernstein trick implies that for all $t \geq 0$,

$$\forall t \geqslant 0, \mathbb{P}\left(\lambda_{max}\left(\mathbf{Z}_n\right) > t\right) \leqslant \inf_{s>0} e^{-st}\mathbb{E}\left[e^{s\lambda_{max}(\mathbf{Z}_n)}\right]. \tag{2.6}$$

Notice that

$$\mathbb{E}\left[e^{s\|\mathbf{Z}_n\|}\right] \leqslant \mathbb{E}\left[e^{s\lambda_{\max}(\mathbf{Z}_n)}\right] + \mathbb{E}\left[e^{s\lambda_{\max}(-\mathbf{Z}_n)}\right] = 2\mathbb{E}\left[e^{s\lambda_{\max}(\mathbf{Z}_n)}\right] \tag{2.7}$$

since $\|\mathbf{Z}_n\| = \max\left\{\lambda_{\max}\left(\mathbf{Z}_n\right), \lambda_{\max}\left(-\mathbf{Z}_n\right)\right\}$ and $\mathbf{Z}_n$ has the same law as $-\mathbf{Z}_n$.

#### 2.2.3.2 Spectral Mapping

The function $x \mapsto e^{sx}$ is monotone, non-decreasing and positive for all $s \geq 0$. It follows from the spectral mapping property (1.60) that for all $s \geq 0$, the largest eigenvalue of $e^{s\mathbf{Z}_n}$ is $e^{s\lambda_{\max}(\mathbf{Z}_n)}$ and all eigenvalues of $e^{s\mathbf{Z}_n}$ are nonnegative. Using the equality "trace $=$ sum of eigenvalues" implies that for all $s \geq 0$,

$$\mathbb{E}\left[e^{s\lambda_{\max}(\mathbf{Z}_n)}\right] = \mathbb{E}\left[\lambda_{\max}\left(e^{s\mathbf{Z}_n}\right)\right] \leqslant \mathbb{E}\left[\mathrm{Tr}\left(e^{s\mathbf{Z}_n}\right)\right]. \tag{2.8}$$

Combining (2.6), (2.7) with (2.8) gives

$$\forall t \geq 0, \mathbb{P}\left(\|\mathbf{Z}_n\| > t\right) \leqslant 2\inf_{s\geq 0} e^{-st}\mathbb{E}\left[\mathrm{Tr}\left(e^{s\mathbf{Z}_n}\right)\right]. \tag{2.9}$$

Up to now, the Oliveira's proof in [38] has followed Ahlswede and Winter's argument in [36]. The next lemma is originally due to Oliveira [38]. Now Oliveira considers the special case

$$\mathbf{Z}_n = \sum_{i=1}^{n} \varepsilon_i \mathbf{A}_i, \tag{2.10}$$

where $\varepsilon_i$ are random coefficients and $\mathbf{A}_1, \ldots, \mathbf{A}_n$ are **deterministic** Hermitian matrices. Recall that a Rademacher sequence is a sequence of $\varepsilon_i{}_{i=1}^{n}$ of i.i.d. random variables with $\varepsilon_i = \varepsilon_1$ uniform over $\{-1, 1\}$. A standard Gaussian sequence is a sequence i.i.d. standard Gaussian random variables.

**Lemma 2.2.6 (Oliveira [38]).** *For all* $s \in \mathbb{R}$*,*

$$\mathbb{E}\left[\mathrm{Tr}\left(e^{s\mathbf{Z}_n}\right)\right] = \mathrm{Tr}\left[\mathbb{E}\left(e^{s\mathbf{Z}_n}\right)\right] \leqslant \mathrm{Tr}\left[\exp\left(\frac{s^2 \sum\limits_{i=1}^{n} \mathbf{A}_i^2}{2}\right)\right]. \tag{2.11}$$

*Proof.* In (2.11), we have used the fact that trace and expectation commute, according to (1.87). The key proof steps have been followed by Rudelson [93], Harvey [91, 92], and Wigderson and Xiao [94]. □

## *2.2.4 Method 4—Ahlswede-Winter's Derivation*

Ahlswede and Winter [36] were the first who used the matrix Laplace transform method. Ahlswede-Winter's derivation, taken from [36], is presented in detail below. We postpone their original version until now, for easy understanding. Their paper and Tropp's long paper [53] are two of the most important sources on this topic. We first digress to study the problem of hypothesis for motivation.

Consider a hypothesis testing problem for a motivation

$$\begin{aligned}\mathcal{H}_0 &: \mathbf{A}_1, \ldots, \mathbf{A}_K \\ \mathcal{H}_1 &: \mathbf{B}_1, \ldots, \mathbf{B}_K\end{aligned}$$

where a sequence of positive, random matrices $\mathbf{A}_i, i = 1, \ldots, K$ and $\mathbf{B}_i, i = 1, \ldots, K$ are considered.

**Algorithm 2.2.7 (Detection Using Traces of Sums of Covariance Matrices).**

1. *Claim* $\mathcal{H}_1$ *if*

$$\mathrm{Tr} \sum_{k=1}^{K} \mathbf{A}_k = \xi \leq \mathrm{Tr} \sum_{k=1}^{K} \mathbf{B}_k,$$

2. *Otherwise, claim* $\mathcal{H}_0$*.*

Only diagonal elements are used in Algorithm 2.2.7; However, non-diagonal elements contain information of use to detection. The exponential of a matrix provides one tool. See Example 2.2.9. In particular, we have

$$\mathrm{Tr}e^{\mathbf{A}+\mathbf{B}} \leq \mathrm{Tr}e^{\mathbf{A}}e^{\mathbf{B}}.$$

The following matrix inequality

$$\mathrm{Tr}e^{\mathbf{A}+\mathbf{B}+\mathbf{C}} \leq \mathrm{Tr}e^{\mathbf{A}}e^{\mathbf{B}}e^{\mathbf{C}}$$

is known to be false.

Let $\mathbf{A}$ and $\mathbf{B}$ be two Hermitian matrices of the same size. If $\mathbf{A} - \mathbf{B}$ is positive semidefinite, we write [16]

$$\mathbf{A} \geq \mathbf{B} \quad \text{or} \quad \mathbf{B} \leq \mathbf{A}. \tag{2.12}$$

$\geq$ is a partial ordering, referred to as Löwner partial ordering, on the set of Hermitian matrices, that is,

1. $\mathbf{A} \geq \mathbf{A}$ for every Hermitian matrix $\mathbf{A}$,
2. If $\mathbf{A} \geq \mathbf{B}$ and $\mathbf{B} \geq \mathbf{A}$, then $\mathbf{A} = \mathbf{B}$, and
3. If $\mathbf{A} \geq \mathbf{B}$ and $\mathbf{B} \geq \mathbf{C}$, then $\mathbf{A} \geq \mathbf{C}$.

The statement $\mathbf{A} \geq 0 \Leftrightarrow \mathbf{X}^*\mathbf{A}\mathbf{X} \geq 0$ is generalized as follows:

$$\mathbf{A} \geq \mathbf{B} \Leftrightarrow \mathbf{X}^*\mathbf{A}\mathbf{X} \geq \mathbf{X}^*\mathbf{B}\mathbf{X} \tag{2.13}$$

for every complex matrix $\mathbf{X}$.

A hypothesis detection problem can be viewed as a problem of partially ordering the measured matrices for individual hypotheses. If many $(K)$ copies of the measured matrices $\mathbf{A}_k$ and $\mathbf{B}_k$ are at our disposal, it is natural to ask this fundamental question:

**Is $\mathbf{B}_1 + \mathbf{B}_2 + \cdots + \mathbf{B}_K$ (statistically) different than $\mathbf{A}_1 + \mathbf{A}_2 + \cdots + \mathbf{A}_K$ ?**

To answer this question motivates this whole section. It turns out that a new theory is needed. We freely use [36] that contains a relatively complete appendix for this topic.

The theory of real random variables provides the framework of much of modern probability theory, such as laws of large numbers, limit theorems, and probability estimates for large deviations, when sums of independent random variables are involved. Researchers develop analogous theories for the case that the algebraic structure of the reals is substituted by more general structures such as groups, vector spaces, etc.

At the hands of our current problem of hypothesis detection, we focus on a structure that has vital interest in quantum probability theory and names the algebra of operators[2] on a (complex) Hilbert space. In particular, the real vector space of self-adjoint operators (Hermitian matrices) can be regarded as a partially ordered generalization of the reals, as reals are embedded in the complex numbers.

#### 2.2.4.1 Fundamentals of Matrix-Valued Random Variables

In the ground-breaking work of [36], they focus on a structure that has vital interest in the algebra of operators on a (complex) Hilbert space, and in particular, the real vector space of self-adjoint operators. Through the spectral mapping theorem, these self-adjoint operators can be regarded as a partially ordered generalization of the reals, as reals are embedded in the complex numbers. To study the convergence of sums of matrix-valued random variables, this partial order is necessary. It will be clear later.

One can generalize the exponentially good estimate for large deviations by the so-called Bernstein trick that gives the famous Chernoff bound [96, 97].

A matrix-valued random variable $\mathbf{X} : \mathbf{\Omega} \to \mathcal{A}_s$, where

$$\mathcal{A}_s = \{\mathbf{A} \in \mathcal{A} : \mathbf{A} = \mathbf{A}^*\} \tag{2.14}$$

is the self-adjoint part of the $\mathbf{C}^*$-algebra $\mathcal{A}$ [98], which is a real vector space. Let $\mathcal{L}(\mathcal{H})$ be the full operator algebra of the complex Hilbert space $\mathcal{H}$. We denote $d = \dim(\mathcal{H})$, which is assumed to be finite. Here $\dim$ means the dimensionality of the vector space. In the general case, $d = \mathrm{Tr}\mathbf{I}$, and $\mathcal{A}$ can be embedded into $\mathcal{L}(\mathcal{C}^d)$ as an algebra, *preserving the trace*. Note the trace (often regarded as expectation) has the property $\mathrm{Tr}\,(\mathbf{AB}) = \mathrm{Tr}\,(\mathbf{BA})$, for any two matrices (or operators) of $\mathbf{A}, \mathbf{B}$. In free probability[3] [99], this is a (optional) axiom as very weak form of commutativity in the trace [9, p. 169].

The real cone

$$\mathcal{A}_+ = \{\mathbf{A} \in \mathcal{A} : \mathbf{A} = \mathbf{A}^* \geq 0\} \tag{2.15}$$

induces a *partial order* $\leq$ in $\mathcal{A}_s$. This partial order is in analogy with the order of two real numbers $a \leq b$. The partial order is the main interest in what follows. We can introduce some convenient notation: for $\mathbf{A}, \mathbf{B} \in \mathcal{A}_s$ the closed interval $[\mathbf{A}, \mathbf{B}]$ is defined as

$$[\mathbf{A}, \mathbf{B}] = \{\mathbf{X} \in \mathcal{A}_s : \mathbf{A} \leq \mathbf{X} \leq \mathbf{B}\}. \tag{2.16}$$

---

[2]The finite-dimensional operators and matrices are used interchangeably.

[3]The idea of free probability is to make algebra (such as operator algebras $C^*$-algebra, von Neumann algebras) the foundation of the theory, as opposed to other possible choices of foundations such as sets, measures, categories, etc.

This is an analogy with the interval $x \in [a, b]$ when $a \leq x \leq b$ for $a, b \in \mathbb{R}$. Similarly, open and half-open intervals $(\mathbf{A}, \mathbf{B})$, $[\mathbf{A}, \mathbf{B})$, etc.

For simplicity, the space $\mathbf{\Omega}$ on which the random variable lives is discrete. Some remarks on the matrix (or operator) order is as follows.

1. The notation "$\leq$" when used for the matrices is *not a total order* unless $\mathcal{A}$, the set of $\mathbf{A}$, spans the entire complex space, i.e., $\mathcal{A} = \mathbb{C}$, in which case the set of self-adjoint operators is the real number space, i.e., $\mathcal{A}_s = \mathbb{R}$. Thus in this case (classical case), the theory developed below reduces to the study of the real random variables.
2. $\mathbf{A} \geq 0$ is equivalent to saying that all eigenvalues of $\mathbf{A}$ are nonnegative. These are $d$ *nonlinear inequalities*. However, we can have the alternative characterization:

$$\begin{aligned} \mathbf{A} \geq 0 &\Leftrightarrow \forall \rho \text{ density operator } \mathrm{Tr}(\rho\mathbf{A}) \geq 0 \\ &\Leftrightarrow \forall \pi \ \text{ one} - \text{dimensional projector } \ \mathrm{Tr}(\pi\mathbf{A}) \geq 0 \end{aligned} \tag{2.17}$$

   From which, we see that these nonlinear inequalities are equivalent to infinitely many *linear* inequalities, which is better adapted to the vector space structure of $\mathcal{A}_s$.
3. The operator mapping $\mathbf{A} \mapsto \mathbf{A}^s$, for $s \in [0, 1]$ and $\mathbf{A} \mapsto \log \mathbf{A}$ are defined on $\mathcal{A}_+$, and both are operator monotone and operator concave. In contrast, $\mathbf{A} \mapsto \mathbf{A}^s$, for $s > 2$ and $\mathbf{A} \mapsto \exp \mathbf{A}$ are neither operator monotone nor operator convex. Remarkably, $\mathbf{A} \mapsto \mathbf{A}^s$, for $s \in [1, 2]$ is operator convex (though not operator monotone). See Sect. 1.4.22 for definitions.
4. The mapping $\mathbf{A} \mapsto \mathrm{Tr} \exp \mathbf{A}$ is monotone and convex. See [50].
5. Golden-Thompson-inequality [23]: for $\mathbf{A}, \mathbf{B} \in \mathcal{A}_s$

$$\mathrm{Tr} \exp(\mathbf{A} + \mathbf{B}) \leq \mathrm{Tr}((\exp \mathbf{A})(\exp \mathbf{B})). \tag{2.18}$$

Items 1–3 follows from Loewner's theorem. A good account of the partial order is [18, 22, 23]. Note that a rarely few of mappings (functions) are operator convex (concave) or operator monotone. Fortunately, we are interested in the trace functions that have much bigger sets [18].

Take a look at (2.19) for example. Since $\mathcal{H}_0 : \mathbf{A} = \mathbf{I} + \mathbf{X}$, and $\mathbf{A} \in \mathcal{A}_s$ (even stronger $\mathbf{A} \in \mathcal{A}_+$), it follows from (2.18) that

$$\mathcal{H}_0 : \mathrm{Tr} \exp(\mathbf{A}) = \mathrm{Tr} \exp(\mathbf{I} + \mathbf{X}) \leq \mathrm{Tr}((\exp \mathbf{I})(\exp \mathbf{X})). \tag{2.19}$$

The use of (2.19) allows us to separately study the diagonal part and the non-diagonal part of the covariance matrix of the noise, since all the diagonal elements are equal for a WSS random process. At low SNR, the goal is to find some ratio or threshold that is statistically stable over a large number of Monte Carlo trials.

**Algorithm 2.2.8 (Ratio detection algorithm using the trace exponentials).**
*1. Claim $\mathcal{H}_1$, if* $\xi = \frac{\mathrm{Tr}\exp\mathbf{A}}{\mathrm{Tr}((\exp\mathbf{I})(\exp\mathbf{X}))} \geq 1$, *where* $\mathbf{A}$ *is the measured covariance matrix with or without signals and* $\mathbf{X} = \frac{\mathbf{R}_w}{\sigma_w^2} - \mathbf{I}$.
*2. Otherwise, claim $\mathcal{H}_0$.*

*Example 2.2.9 (Exponential of the $2 \times 2$ matrix).* The $2 \times 2$ covariance matrix for $L$ sinusoidal signals has symmetric structure with identical diagonal elements

$$\mathbf{R}_s = \mathrm{Tr}\mathbf{R}_s(\mathbf{I} + b\sigma_1)$$

where

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

and $b$ is a positive number. Obviously, $\mathrm{Tr}\sigma_1 = 0$. We can study the diagonal elements and non-diagonal elements separately. The two eigenvalues of the $2 \times 2$ matrix [100]

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

are

$$\lambda_{1,2} = \tfrac{1}{2}\mathrm{Tr}\mathbf{A} \pm \tfrac{1}{2}\sqrt{\mathrm{Tr}^2\mathbf{A} - 4\det\mathbf{A}}$$

and the corresponding eigenvectors are, respectively,

$$u_1 = \frac{1}{||u_1||}\begin{pmatrix} b \\ \lambda_1 - a \end{pmatrix}; \quad u_2 = \frac{1}{||u_2||}\begin{pmatrix} b \\ \lambda_2 - a \end{pmatrix}.$$

To study how the zero-trace $2 \times 2$ matrix $\sigma_1$ affects the exponential, consider

$$\mathbf{X} = \begin{pmatrix} 0 & b \\ a^{-1} & 0 \end{pmatrix}.$$

The exponential of the matrix $\mathbf{X}$, $e^{\mathbf{X}}$, has positive entries, and in fact [101]

$$e^{\mathbf{X}} = \begin{pmatrix} \cosh\sqrt{\frac{b}{a}} & \sqrt{ab}\sinh\sqrt{\frac{b}{a}} \\ \frac{1}{\sqrt{ab}}\sinh\sqrt{\frac{b}{a}} & \cosh\sqrt{\frac{b}{a}} \end{pmatrix}$$

□

#### 2.2.4.2 Matrix-Valued Concentration Inequalities

In analogy with the scalar-valued random variables, we can develop a matrix-valued Markov inequality. Suppose that $X$ is a nonnegative random variable with mean $\mathbb{E}[X]$. The scalar-valued **Markov inequality** of (1.6) states that

$$\mathbb{P}(X \geqslant a) \leqslant \frac{\mathbb{E}[X]}{a} \quad \text{for } X \text{ nonnegative.} \tag{2.20}$$

**Theorem 2.2.10 (Markov inequality).** *Let* $\mathbf{X}$ *a matrix-valued random variable with values in* $\mathcal{A}_+$ *and expectation*

$$\mathbf{M} = \mathbb{E}\mathbf{X} = \sum_{\mathcal{X}} \Pr\{\mathbf{X} = \mathcal{X}\}\mathcal{X}, \tag{2.21}$$

*and* $\mathbf{A} \geq 0$ *is a fixed positive semidefinite matrix. Then*

$$\Pr\{\mathbf{X} \nleq \mathbf{A}\} \leq \mathrm{Tr}\left(\mathbf{M}\mathbf{A}^{-1}\right). \tag{2.22}$$

*Proof.* The support of $\mathbf{A}$ is assumed to contain the support of $\mathbf{M}$, otherwise, the theorem is trivial. Let us consider the positive matrix-valued random variable $\mathbf{Y} = \mathbf{A}^{-1/2}\mathbf{X}\mathbf{A}^{-1/2}$ which has expectation $\mathbb{E}[\mathbf{Y}] = \mathbf{A}^{-1/2}\mathbb{E}[\mathbf{X}]\mathbf{A}^{-1/2}$, using the product rule of (1.88): $\mathbb{E}[\mathbf{XY}] = \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{Y}]$. we have used the fact that the expectation of a constant matrix is itself: $\mathbb{E}[\mathbf{A}] = \mathbf{A}$.

Since the events $\{\mathbf{X} \leqslant \mathbf{A}\}$ and $\{\mathbf{Y} \leqslant \mathbf{I}\}$ coincide, we have to show that

$$\mathbb{P}(\mathbf{Y} \leqslant \mathbf{I}) \nleqslant \mathrm{Tr}(\mathbb{E}[\mathbf{X}])$$

Note from (1.87), the trace and expectation commute! This is seen as follows:

$$\mathbb{E}[\mathbf{Y}] = \sum_{\mathcal{Y}} \mathbb{P}(\mathbf{Y} = \mathcal{Y})\,\mathcal{Y} \geqslant \sum_{\mathcal{Y} \nleq \mathbf{I}} \mathbb{P}(\mathbf{Y} = \mathcal{Y})\,\mathcal{Y}.$$

The second inequality follows from the fact that $\mathbf{Y}$ is positive and $\mathcal{Y} = \{\mathcal{Y} \nleqslant \mathbf{I}\} \cup \{\mathcal{Y} > \mathbf{I}\}$. All eigenvalues of $\mathbf{Y}$ are positive. Ignoring the event $\{\mathcal{Y} > \mathbf{I}\}$ is equivalent to remove some positive eigenvalues from the spectrum of the $\mathbf{Y}$, $\mathrm{spec}(\mathbf{Y})$.

Taking traces, and observing that a positive operator (or matrix) which is not less than or equal to $\mathbf{I}$ must have trace at least 1, we find

$$\mathrm{Tr}(\mathbb{E}[\mathbf{Y}]) \geqslant \sum_{\mathcal{Y} \nleq \mathbf{I}} \mathbb{P}(\mathbf{Y} = \mathcal{Y})\,\mathrm{Tr}(\mathcal{Y}) \geqslant \sum_{\mathcal{Y} \nleq \mathbf{I}} \mathbb{P}(\mathbf{Y} = \mathcal{Y}) = \mathbb{P}(\mathbf{Y} \nleq \mathbf{I}),$$

which is what we wanted. □

In the case of $\mathcal{H} = \mathbb{C}$ the theorem reduces to the well-known Markov inequality for nonnegative real random variables. One easily see that, as in the classical case, the inequality is optimal in the sense that there are examples when the inequality is assumed with equality.

Suppose that the mean $m$ and the variance $\sigma^2$ of a scalar-valued random variable $X$ are known. The *Chebyshev inequality* of (1.8) states that

$$\mathbb{P}\left(|X - m| \geqslant a\right) \leqslant \frac{\sigma^2}{a^2}. \tag{2.23}$$

The Chebyshev inequality is a consequence of the Markov inequality.

In analogy with the scalar case, if we assume knowledge about the matrix-valued expectation and the matrix-valued variance, we can prove the matrix-valued Chebyshev inequality.

**Theorem 2.2.11 (Chebyshev inequality).** *Let* $\mathbf{X}$ *a matrix-valued random variable with values in* $\mathcal{A}_s$*, expectation* $\mathbf{M} = \mathbb{E}\mathbf{X}$*, and variance*

$$\mathrm{Var}\mathbf{X} = \mathbf{S}^2 = \mathbb{E}\left((\mathbf{X} - \mathbf{M})^2\right) = \mathbb{E}(\mathbf{X}^2) - \mathbf{M}^2. \tag{2.24}$$

*For* $\boldsymbol{\Delta} \geq 0$*,*

$$\mathbb{P}\{|\mathbf{X} - \mathbf{M}| \nleq \boldsymbol{\Delta}\} \leq \mathrm{Tr}\left(\mathbf{S}^2\boldsymbol{\Delta}^{-2}\right). \tag{2.25}$$

*Proof.* Observe that

$$|\mathbf{X} - \mathbf{M}| \leq \boldsymbol{\Delta} \Leftarrow (\mathbf{X} - \mathbf{M})^2 \leq \boldsymbol{\Delta}^2$$

since $\sqrt{(\cdot)}$ is operator monotone. See Sect. 1.4.22. We find that

$$\mathbb{P}\left(|\mathbf{X} - \mathbf{M}| \nleq \boldsymbol{\Delta}\right) \leqslant \mathbb{P}\left((\mathbf{X} - \mathbf{M})^2 \nleq \boldsymbol{\Delta}^2\right) \leqslant \mathrm{Tr}\left(\mathbf{S}^2\boldsymbol{\Delta}^{-2}\right).$$

The last step follows from Theorem 2.2.10. □

If $\mathbf{X}, \mathbf{Y}$ are independent, then $\mathrm{Var}(\mathbf{X} + \mathbf{Y}) = \mathrm{Var}\mathbf{X} + \mathrm{Var}\mathbf{Y}$. This is the same as in the classical case but one has to pay attention to the noncommunicativity that causes technical difficulty.

**Corollary 2.2.12 (Weak law of large numbers).** *Let* $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ *be identically, independently, distributed (i.i.d.) matrix-valued random variables with values in* $\mathcal{A}_s$*, expectation* $\mathbf{M} = \mathbb{E}\mathbf{X}$*, and variance* $\mathrm{Var}\mathbf{X} = \mathbf{S}^2$*. For* $\boldsymbol{\Delta} \geq 0$*, then*

$$\begin{aligned}
&\mathbb{P}\left\{\frac{1}{n}\sum_{n=1}^{n}\mathbf{X}_i \notin [\mathbf{M}-\mathbf{\Delta},\mathbf{M}+\mathbf{\Delta}]\right\} \le \frac{1}{n}\mathrm{Tr}\left(\mathbf{S}^2\mathbf{\Delta}^{-2}\right),\\
&\mathbb{P}\left\{\sum_{n=1}^{n}\mathbf{X}_i \notin [n\mathbf{M}-\sqrt{n}\mathbf{\Delta},n\mathbf{M}-\sqrt{n}\mathbf{\Delta}]\right\} \le \frac{1}{n}\mathrm{Tr}\left(\mathbf{S}^2\mathbf{\Delta}^{-2}\right).
\end{aligned} \tag{2.26}$$

*Proof.* Observe that $\mathbf{Y} \notin [\mathbf{M}-\mathbf{\Delta},\mathbf{M}+\mathbf{\Delta}]$ is equivalent to $|\mathbf{Y}-\mathbf{M}| \nleq \mathbf{\Delta}$, and apply the previous theorem. The event $|\mathbf{Y}-\mathbf{M}| \nleq \mathbf{\Delta}$ says that the absolute values of the eigenvalues of the matrix $\mathbf{Y}-\mathbf{M}$ is bounded by the eigenvalues of $\mathbf{\Delta}$ which are of course nonnegative. The matrix $\mathbf{Y}-\mathbf{M}$ is Hermitian (but not necessarily nonnegative or nonpositive). That is why the absolute value operation is needed. □

When we see these functions of matrix-valued inequalities, we see the functions of their eigenvalues. The spectral mapping theorem must be used all the time.

**Lemma 2.2.13 (Large deviations and Bernstein trick).** *For a matrix-valued random variable* $\mathbf{Y}$, $\mathbf{B} \in \mathcal{A}_s$, *and* $\mathbf{T} \in \mathcal{A}$ *such that* $\mathbf{T}^*\mathbf{T} > 0$

$$\mathbb{P}\left\{\mathbf{Y} \nleq \mathbf{B}\right\} \le \mathrm{Tr}\left[\mathbb{E}\mathrm{e}^{\mathbf{TYT}^* - \mathbf{TBT}^*}\right]. \tag{2.27}$$

*Proof.* We directly calculate

$$\begin{aligned}
\mathbb{P}\left(\mathbf{Y} \nleq \mathbf{B}\right) &= \mathbb{P}\left(\mathbf{Y}-\mathbf{B} \nleq 0\right)\\
&= \mathbb{P}\left(\mathbf{TYT}^* - \mathbf{TBT}^* \nleq 0\right)\\
&= \mathbb{P}\left[\mathrm{e}^{\mathbf{TYT}^* - \mathbf{TBT}^*} \nleq \mathbf{I}\right]\\
&\leqslant \mathrm{Tr}\left[\mathbb{E}\mathrm{e}^{\mathbf{TYT}^* - \mathbf{TBT}^*}\right].
\end{aligned}$$

Here, the second line is because the mapping $\mathbf{X} \mapsto \mathbf{TXT}^*$ is bijective and preserve the order. The $\mathbf{TYT}^*$ and $\mathbf{TBT}^*$ are two *commutative* matrices. For commutative matrices $\mathbf{A},\mathbf{B}$, $\mathbf{A} \le \mathbf{B}$ is equivalent to $e^{\mathbf{A}} \le e^{\mathbf{B}}$, from which the third line follows. The last line follows from Theorem 2.2.10. □

The Bernstein trick is a crucial step. The problem is reduced to the form of $\mathrm{Tr}\left[\mathbb{E}\mathrm{e}^{\mathbf{Z}}\right]$ where $\mathbf{Z} = \mathbf{TYT}^* - \mathbf{TBT}^*$ is Hermitian. We really do not know if $\mathbf{Z}$ is nonnegative or positive. But we do not care since the matrix exponential of any Hermitian $\mathbf{A}$ is always nonnegative. As a consequence of using the Bernstein trick, we only need to deal with nonnegative matrices.

But we need another key ingredient—Golden-Thompson inequality—since for Hermitian $\mathbf{AB}$, we have $e^{\mathbf{A}+\mathbf{B}} \ne e^{\mathbf{A}} \cdot e^{\mathbf{B}}$, unlike $e^{a+b} = e^a \cdot e^b$, for two scalars $a, b$. For two Hermitian matrices $\mathbf{A},\mathbf{B}$, we have the Golden-Thompson inequality

$$\mathrm{Tr}\left(e^{\mathbf{A}+\mathbf{B}}\right) \leqslant \mathrm{Tr}\left(e^{\mathbf{A}} \cdot e^{\mathbf{B}}\right).$$

**Theorem 2.2.14 (Concentration for i.i.d matrix-valued random variables).** *Let $\mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n$ be i.i.d. matrix-valued random variables with values in $\mathcal{A}_s$, $\mathbf{A} \in \mathcal{A}_s$. Then for $\mathbf{T} \in \mathcal{A}$, $\mathbf{T}^*\mathbf{T} > 0$*

$$\mathbb{P}\left\{ \sum_{n=1}^{n} \mathbf{X}_i \nleqslant n\mathbf{A} \right\} \leqslant d \cdot \mathbb{E}\|\exp\left(\mathbf{TXT}^* - \mathbf{TAT}^*\right)\|^n. \tag{2.28}$$

Define the binary I-divergence as

$$D(u\|v) = u(\log u - \log v) + (1-u)(\log(1-u) - \log(1-v)). \tag{2.29}$$

*Proof.* Using previous lemma (obtained from the Bernstein trick) with $\mathbf{Y} = \sum_{i=1}^{n} \mathbf{X}_i$ and $\mathbf{B} = n\mathbf{A}$, we find

$$\begin{aligned}
\mathbb{P}\left\{ \sum_{i=1}^{n} \mathbf{X}_i \nleqslant n\mathbf{A} \right\} &\leqslant \operatorname{Tr}\left\{ \mathbb{E}\exp\left[ \sum_{i=1}^{n} \mathbf{T}\left(\mathbf{X}_i - \mathbf{A}\right)\mathbf{T}^* \right] \right\} \\
&= \mathbb{E}\left\{ \operatorname{Tr}\exp\left[ \sum_{i=1}^{n} \mathbf{T}\left(\mathbf{X}_i - \mathbf{A}\right)\mathbf{T}^* \right] \right\} \\
&\leqslant \mathbb{E}\operatorname{Tr}\left\{ \exp\left[ \sum_{i=1}^{n-1} \mathbf{T}\left(\mathbf{X}_i - \mathbf{A}\right)\mathbf{T}^* \right] \exp\left[\mathbf{T}\left(\mathbf{X}_n - \mathbf{A}\right)\mathbf{T}^*\right] \right\} \\
&= \mathbb{E}_{\mathbf{X}_1,\ldots,\mathbf{X}_{n-1}} \operatorname{Tr}\left\{ \exp\left[ \sum_{i=1}^{n-1} \mathbf{T}\left(\mathbf{X}_i - \mathbf{A}\right)\mathbf{T}^* \right] \mathbb{E}\exp\left[\mathbf{T}\left(\mathbf{X}_n - \mathbf{A}\right)\mathbf{T}^*\right] \right\} \\
&\leqslant \|\mathbb{E}\exp\left[\mathbf{T}\left(\mathbf{X}_n - \mathbf{A}\right)\mathbf{T}^*\right]\| \cdot \mathbb{E}_{\mathbf{X}_1,\ldots,\mathbf{X}_{n-1}} \operatorname{Tr}\left\{ \exp\left[ \sum_{i=1}^{n-1} \mathbf{T}\left(\mathbf{X}_i - \mathbf{A}\right)\mathbf{T}^* \right] \right\} \\
&\leqslant d \cdot \|\mathbb{E}\exp\left[\mathbf{T}\left(\mathbf{X}_n - \mathbf{A}\right)\mathbf{T}^*\right]\|^n
\end{aligned}$$

the first line follows from Lemma 2.2.13. The second line is from the fact that the trace and expectation commute, according to (1.87). In the third line, we use the famous Golden-Thompson inequality (1.71). In the fourth line, we take the expectation on the $\mathbf{X}_n$. The fifth line is due to the norm property (1.84). The sixth line is using the fifth step by induction for $n$ times. $d$ comes from the fact $\operatorname{Tr}\exp\left(\mathbf{0}\right) = \operatorname{Tr}\mathbf{I} = d$, where $\mathbf{0}$ is a zero matrix whose entries are all zero. □

The problem is now to minimize $\|\mathbb{E}\exp\left[\mathbf{T}\left(\mathbf{X}_n - \mathbf{A}\right)\mathbf{T}^*\right]\|$ with respect to $\mathbf{T}$. For a Hermitian matrix $T$, we have the polar decomposition $\mathbf{T} = |\mathbf{T}| \cdot \mathbf{U}$, where $\mathbf{Y}$ is a unitary matrix; so, without loss of generality, we may assume that $\mathbf{T}$ is Hermitian. Let us focus pursue the special case of a bounded matrix-valued random variable. Defining

$$D\left(u\|v\right) = u\left(\log u - \log v\right) + \left(1-u\right)\left[\log\left(1-u\right) - \log\left(1-v\right)\right]$$

we find the following matrix-valued Chernoff bound.

**Theorem 2.2.15 (Matrix-valued Chernoff Bound).** *Let* $\mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n$ *be i.i.d. matrix-valued random variables with values in* $[\mathbf{0}, \mathbf{I}] \in \mathcal{A}_s$ , $\mathbb{E}\mathbf{X} \leq m\mathbf{I}$*,* $\mathbf{A} \geq a\mathbf{I}$, $1 \geq a \geq m \geq 0$. *Then*

$$\mathbb{P}\left\{\sum_{n=1}^{n} \mathbf{X}_i \nleq n\mathbf{A}\right\} \leq d \cdot \exp\left(-nD\left(a||m\right)\right), \tag{2.30}$$

*Similarly,* $\mathbb{E}\mathbf{X} \geq m\mathbf{I}$*,* $\mathbf{A} \leq a\mathbf{I}$, $0 \leq a \leq m \leq 1$. *Then*

$$\mathbb{P}\left(\sum_{n=1}^{n} \mathbf{X}_i \ngeq n\mathbf{A}\right) \leqslant d \cdot \exp\left(-nD\left(a||m\right)\right), \tag{2.31}$$

*As a consequence, we get, for* $\mathbb{E}\mathbf{X} = \mathbf{M} \geq \mu\mathbf{I}$ *and* $0 \leq \epsilon \leq \frac{1}{2}$*, then*

$$\mathbb{P}\left\{\frac{1}{n}\sum_{n=1}^{n} \mathbf{X}_i \notin \left[\left(1-\varepsilon\right)\mathbf{M}, \left(1+\varepsilon\right)\mathbf{M}\right]\right\} \leq 2d \cdot \exp\left(-n \cdot \frac{\varepsilon^2\mu}{2\ln 2}\right). \tag{2.32}$$

*Proof.* The second part follows from the first by considering $\mathbf{Y}_i = \mathbf{X}_i$, and the observation that $D\left(a||m\right) = D\left(1-a||1-m\right)$. To prove it, we apply Theorem 2.2.14 with a special case of $\mathbf{T} = \sqrt{t}\mathbf{I}$ to obtain

$$\begin{aligned}\mathbb{P}\left\{\sum_{n=1}^{n} \mathbf{X}_i \nleq n\mathbf{A}\right\} &\leqslant \mathbb{P}\left\{\sum_{n=1}^{n} \mathbf{X}_i \nleq na\mathbf{I}\right\}\\ &\leqslant d \cdot ||\mathbb{E}\exp(t\mathbf{X})\exp(-ta\mathbf{I})||^n\\ &= d \cdot ||\mathbb{E}\exp(t\mathbf{X})\exp(-ta)||^n.\end{aligned}$$

Note $\exp(-ta\mathbf{I}) = \exp(-ta)\mathbf{I}$ and $\mathbf{AI} = \mathbf{A}$. Now the convexity of the exponential function $\exp(x)$ implies that

$$\frac{\exp\left(tx\right)-1}{x} \leqslant \frac{\exp\left(t\right)-1}{1}, 0 \leqslant x \leqslant 1, x \in \mathbb{R},$$

which, by replacing $x$ with matrix $\mathbf{X} \in \mathcal{A}_\mathbf{s}$ and $1$ with the identify matrix $\mathbf{I}$ (see Sect. 1.4.13 for this rule), yields

$$\exp(t\mathbf{X}) - 1 \leqslant \mathbf{X}\left(\exp\left(t\right)-1\right).$$

As a consequence, we have

$$\mathbb{E}\exp(t\mathbf{X}) \leqslant \mathbf{I} + (\exp(t) - 1)\,\mathbb{E}\mathbf{X}. \tag{2.33}$$

hence, we have

$$\begin{aligned}
||\mathbb{E}\exp(t\mathbf{X})\exp(-ta)|| &\leqslant \exp(-ta)\,\|\mathbf{I} + (\exp(t) - 1)\,\mathbb{E}\mathbf{X}\| \\
&\leq \exp(-ta)\,\|\mathbf{I} + (\exp(t) - 1)\,m\mathbf{I}\| \\
&= \exp(-ta)\,[1 + (\exp(t) - 1)\,m]\,.
\end{aligned}$$

The first line follows from using (2.33). The second line follows from the hypothesis of $\mathbb{E}\mathbf{X} \leqslant m\mathbf{I}$. The third line follows from using the spectral norm property of (1.57) for the identity matrix $\mathbf{I}$:$\|\mathbf{I}\| = 1$. Choosing

$$t = \log\left(\frac{a}{m}\cdot\frac{1-m}{1-a}\right) > 0$$

the right-hand side becomes exactly $\exp\left(-D\left(a||m\right)\right)$.

To prove the last claim of the theorem, consider the variables $\mathbf{Y}_i = \mu\mathbf{M}^{-1/2}\mathbf{X}_i\mathbf{M}^{-1/2}$ with expectation $\mathbb{E}\mathbf{Y}_i = \mu\mathbf{I}$ and $\mathbf{Y}_i \in [\mathbf{0}, \mathbf{I}]$, by hypothesis. Since

$$\frac{1}{n}\sum_{n=1}^{n}\mathbf{X}_i \in [(1-\varepsilon)\,\mathbf{M}, (1+\varepsilon)\,\mathbf{M}] \Leftrightarrow \frac{1}{n}\sum_{n=1}^{n}\mathbf{Y}_i \in [(1-\varepsilon)\,\mu\mathbf{I}, (1+\varepsilon)\,\mu\mathbf{I}]$$

we can apply what we just proved to obtain

$$\begin{aligned}
&\mathbb{P}\left\{\frac{1}{n}\sum_{n=1}^{n}\mathbf{X}_i \notin [(1-\varepsilon)\,\mathbf{M}, (1+\varepsilon)\,\mathbf{M}]\right\} \\
&= \mathbb{P}\left\{\frac{1}{n}\sum_{n=1}^{n}\mathbf{X}_i \geqslant (1+\varepsilon)\,\mathbf{M}\right\} + \mathbb{P}\left\{\frac{1}{n}\sum_{n=1}^{n}\mathbf{X}_i \leqslant (1-\varepsilon)\,\mathbf{M}\right\} \\
&\leqslant d\left\{\exp\left[-nD\left((1-\varepsilon)\,\mu||\mu\right)\right] + \exp\left[-nD\left((1+\varepsilon)\,\mu||\mu\right)\right]\right\} \\
&\leqslant 2d\cdot\exp\left(-n\frac{\varepsilon^2\mu}{2\ln 2}\right).
\end{aligned}$$

The last line follows from the already used inequality

$$D\left((1+x)\mu||\mu\right) \geqslant \frac{x^2\mu}{2\ln 2}.$$

$\square$

### 2.2.5 Derivation Method 5—Gross, Liu, Flammia, Becker, and Eisert

Let $||\mathbf{A}||$ be the operator norm of matrix $\mathbf{A}$.

**Theorem 2.2.16 (Matrix-Bernstein inequality—Gross [102]).** *let* $\mathbf{X}_i, i = 1, \ldots, N$ *be i.i.d. zero mean, Hermitian matrix-valued random variables of size* $n \times n$. *Assume* $\sigma_0, c \in \mathbb{R}$ *are such that* $\left\|\mathbf{X}_i^2\right\| \leqslant \sigma_0^2$ *and* $\|\mathbf{X}_i\| \leqslant \mu$. *Set* $\mathbf{S} = \sum\limits_{i=1}^{N} \mathbf{X}_i$ *and let* $\sigma^2 = N\sigma_0^2$, *an upper bound to the variance of* $\mathbf{S}$. *Then*

$$
\begin{aligned}
\mathbb{P}\left(\|\mathbf{S}\| > t\right) &\leqslant 2n \exp\left(-\frac{t^2}{4\sigma^2}\right), \quad t \leqslant \frac{2\sigma}{\mu}, \\
\mathbb{P}\left(\|\mathbf{S}\| < t\right) &\leqslant 2n \exp\left(-\frac{t}{2\mu}\right), \quad t > \frac{2\sigma}{\mu}.
\end{aligned}
\tag{2.34}
$$

We refer to [102] for a proof. His proof directly follows from Ahlswede-Winter [36] with some revisions.

### 2.2.6 Method 6—Recht's Derivation

The version of derivation, taken from [103], is more general in that the random matrices need not be identically distributed. A symmetric matrix is assumed. It is our conjecture that results of [103] may be easily extended to a Hermitian matrix.

**Theorem 2.2.17 (Noncommutative Bernstein Inequality [104]).** *Let* $\mathbf{X}_1, \ldots, \mathbf{X}_L$ *be independent zero-mean random matrices of dimension* $d_1 \times d_2$. *Suppose* $\rho_k^2 = \max\left\{\left\|\mathbb{E}\left(\mathbf{X}_k\mathbf{X}_k^*\right)\right\|, \left\|\mathbb{E}\left(\mathbf{X}_k^*\mathbf{X}_k\right)\right\|\right\}$ *and* $\|\mathbf{X}_k\| \leqslant M$ *almost surely for all* $k$. *Then, for any* $\tau > 0$,

$$
\mathbb{P}\left[\left\|\sum_{k=1}^{L} \mathbf{X}_k\right\| > \tau\right] \leqslant (d_1 + d_2) \exp\left(\frac{-\frac{\tau^2}{2}}{\sum\limits_{k=1}^{L} \rho_k^2 + M\tau/3}\right). \tag{2.35}
$$

Note that in the case that $d_1 = d_2 = 1$, this is precisely the two sided version of the standard Bernstein Inequality. When the $\mathbf{X}_k$ are diagonal, this bound is the same as applying the standard Bernstein Inequality and a union bound to the diagonal of the matrix summation. Besides, observe that the right hand side is less than $(d_1 + d_2) \exp\left(\frac{-\frac{3}{8}\tau^2}{\sum\limits_{k=1}^{L} \rho_k^2}\right)$ as long as $\tau \leqslant \frac{1}{M} \sum\limits_{k=1}^{L} \rho_k^2$. This condensed form of the inequality is used exclusively throughout in [103]. Theorem 2.2.17 is a corollary of an Chernoff bound for finite dimensional operators developed by Ahlswede and Winter [36]. A similar inequality for symmetric i.i.d. matrices is proposed in [95] $||\cdot||$ denotes the spectral norm (the top singular value) of an operator.

### *2.2.7 Method 7—Derivation by Wigderson and Xiao*

Chernoff bounds are extremely useful in probability. Intuitively, they say that a random sample approximates the average, with a probability of deviation that goes down exponentially with the number of samples. Typically we are concerned about real-valued random variables, but recently several applications have called for large-deviations bounds for matrix-valued random variables. Such a bound was given by Ahlswede and Winter [36, 105].

All of Wigderson and Xiao's results [94] are extended to complex Hermitian matrices, or abstractly to self-adjoint operators over any Hilbert spaces where the operations of addition, multiplication, trace exponential, and norm are efficiently computable. Wigderson and Xiao [94] essentially follows the original style of Ahlswede and Winter [36] in the validity of their method.

### *2.2.8 Method 8—Tropp's Derivation*

The derivation follows [53].

$$\begin{aligned}
&\mathbb{P}\left(\left\|\lambda_{\max}\left(\sum_{i=1}^{n}\mathbf{X}_i\right)\right\|\geqslant t\right)\\
&=\mathbb{P}\left(\left\|\lambda_{\max}\left(\sum_{i=1}^{n}\theta\mathbf{X}_i\right)\right\|\geqslant e^{\theta t}\right)\text{(the positive homogeneity of the eigenvalue map)}\\
&\leqslant e^{-\theta t}\cdot\mathbb{E}\exp\left\{\lambda_{\max}\left(\sum_{i=1}^{n}\theta\mathbf{X}_i\right)\right\}\text{(Markov's inequality)}\\
&=e^{-\theta t}\cdot\mathbb{E}\lambda_{\max}\left(\exp\left\{\sum_{i=1}^{n}\theta\mathbf{X}_i\right\}\right)\text{(the spectral mapping theorem)}\\
&<e^{-\theta t}\cdot\mathbb{E}\,\mathrm{Tr}\left(\exp\left\{\sum_{i=1}^{n}\theta\mathbf{X}_i\right\}\right)\text{(the exponential of a Hermitian matrix is positive definite)}
\end{aligned} \tag{2.36}$$

## 2.3 Cumulate-Based Matrix-Valued Laplace Transform Method

This section develops some general probability inequalities for the maximum eigenvalue of a sum of independent random matrices. The main ingredient is a matrix extension of the scalar-valued Laplace transform method for sums of independent real random variables, see Sect. 1.1.6.

Before introducing the matrix-valued Laplace transform, we need to define matrix and cumulants, in analogy with Sect. 1.1.6 for the scalar setting. At this point, a quick review of Sect. 1.1 will illuminate the contrast between the scalar and matrix settings. The central idea of Ahswede and Winter [36] is to extend the textbook idea of the Laplace Transform Method from the scalar setting to the matrix setting.

Consider a Hermitian matrix $\mathbf{X}$ that has moments of all orders. By analogy with the classical scalar definitions (Sect. 1.1.7), we may construct matrix extensions of the moment generating function and the cumulant generating function:

$$\mathbf{M}_{\mathbf{X}}(\theta) := \mathbb{E}e^{\theta\mathbf{X}} \quad \text{and} \quad \mathbf{\Xi}_{\mathbf{X}}(\theta) := \log \mathbb{E}e^{\theta\mathbf{X}} \quad \text{for} \quad \theta \in \mathbb{R}. \tag{2.37}$$

We have the formal power series expansions:

$$\mathbf{M}_{\mathbf{X}}(\theta) = \mathbf{I} + \sum_{k=1}^{\infty} \frac{\theta^k}{k!} \cdot \left(\mathbf{X}^k\right) \quad \text{and} \quad \Xi_{\mathbf{X}}(\theta) = \sum_{k=1}^{\infty} \frac{\theta^k}{k!} \cdot \mathbf{\Xi}_k.$$

The coefficients $\left(\mathbb{E}\mathbf{X}^k\right)$ are called matrix moments and $\mathbf{\Xi}_k$ are called matrix cumulants. The matrix cumulant $\mathbf{\Xi}_k$ has a formal expression as a noncommutative polynomial in the matrix moments up to order $k$. In particular, the first cumulant is the mean and the second cumulant is the variance:

$$\mathbf{\Xi}_1 = \mathbb{E}(\mathbf{X}) \quad \text{and} \quad \mathbf{\Xi}_2 = \mathbb{E}\left(\mathbf{X^2}\right) - \mathbb{E}(\mathbf{X})^2.$$

Higher-order cumulants are harder to write down and interpret.

**Proposition 2.3.1 (The Lapalce Transform Method).** *Let $\mathbf{Y}$ be a random Hermitian matrix. For all $t \in \mathbb{R}$,*

$$\mathbb{P}\left(\lambda_{\max}(\mathbf{Y}) \geqslant t\right) \leqslant \inf_{\theta > 0} \left\{ e^{-\theta t} \cdot \mathbb{E}\,\mathrm{Tr}\, e^{\theta\mathbf{Y}} \right\}.$$

In words, we can control tail probabilities for the maximum eigenvalue of a random matrix by producing a bound for the trace of the matrix moment generating function defined in (2.37). Let us show how Bernstein's Laplace transform technique extends to the matrix setting. The basic idea is due to Ahswede-Winter [36], but we follow Oliveria [38] in this presentation.

*Proof.* Fix a positive number $\theta$. Observe that

$$\mathbb{P}\left(\lambda_{\max}\left(\mathbf{Y}\right)\geqslant t\right)=\mathbb{P}\left(\lambda_{\max}\left(\theta\mathbf{Y}\right)\geqslant\theta t\right)=\mathbb{P}\left(e^{\lambda_{\max}(\theta\mathbf{Y})}\geqslant e^{\theta t}\right)\leqslant e^{-\theta t}\cdot\mathbb{E}e^{\lambda_{\max}(\theta\mathbf{Y})}.$$

The first identity uses the homogeneity of the maximum eigenvalue map. The second relies on the monotonicity of the scalar exponential functions; the third relation is Markov's inequality. To bound the exponential, note that

$$e^{\lambda_{\max}(\theta\mathbf{Y})}=\lambda_{\max}\left(e^{\theta\mathbf{Y}}\right)\leqslant\operatorname{Tr}e^{\theta\mathbf{Y}}.$$

The first relation is the spectral mapping theorem (Sect. 1.4.13). The second relation holds because the exponential of an Hermitian matrix is always positive definite—the eigenvalues of the matrix exponential are always positive (see Sect. 1.4.16 for the matrix exponential); thus, the maximum eigenvalue of a positive definite matrix is dominated by the trace. Combine the latter two relations, we reach

$$\mathbb{P}\left(\lambda_{\max}\left(\mathbf{Y}\right)\geqslant t\right)\leqslant\inf_{\theta>0}\left\{e^{-\theta t}\cdot\mathbb{E}\operatorname{Tr}e^{\theta\mathbf{Y}}\right\}.$$

This inequality holds for any positive $\theta$, so we may take an infimum[4] to complete the proof. □

## 2.4 The Failure of the Matrix Generating Function

In the scalar setting of Sect. 1.2, the Laplace transform method is very effective for studying sums of independent (scalar-valued) random variables, because the matrix generating function decomposes. Consider an independent sequence $X_k$ of real random variables. Operating formally, we see that the scalar matrix generating function of the sum satisfies a multiplicative rule:

$$M_{\left(\sum\limits_k X_k\right)}(\theta)=\mathbb{E}\exp\left(\sum_k\theta X_k\right)=\mathbb{E}\prod_k e^{\theta X_k}=\prod_k\mathbb{E}e^{\theta X_k}=\prod_k M_{X_k}(\theta). \tag{2.38}$$

This calculation relies on the fact that the scalar exponential function converts sums to products, a property the matrix exponential does not share, see Sect. 1.4.16. Thus, there is no immediate analog of (2.38) in the matrix setting.

---

[4]In analysis the infimum or greatest lower bound of a subset $S$ of real numbers is denoted by $\inf(S)$ and is defined to be the biggest real number that is smaller than or equal to every number in $S$. An important property of the real numbers is that every set of real numbers has an infimum (any bounded nonempty subset of the real numbers has an infimum in the non-extended real numbers). For example, $\inf\{1,2,3\}=1, \inf\{x\in\mathbb{R},0<x<1\}=0$.

Ahlswede and Winter attempts to imitate the multiplicative rule of (2.38) using the following observation. When $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent random matrices,

$$\operatorname{Tr}\mathbf{M}_{\mathbf{X}_1+\mathbf{X}_2}(\theta) \leqslant \mathbb{E}\operatorname{Tr}\left[e^{\theta\mathbf{X}_1}e^{\theta\mathbf{X}_2}\right] = \operatorname{Tr}\left[\left(\mathbb{E}e^{\theta\mathbf{X}_1}\right)\left(\mathbb{E}e^{\theta\mathbf{X}_2}\right)\right] = \operatorname{Tr}\left[\mathbf{M}_{\mathbf{X}_1}(\theta)\cdot\mathbf{M}_{\mathbf{X}_2}(\theta)\right]. \tag{2.39}$$

The first relation is the Golden-Thompson trace inequality (1.71). Unfortunately, we cannot extend the bound (2.39) to include additional matrices. This cold fact may suggest that the Golden-Thompson inequality may not be the natural way to proceed. In Sect. 2.2.4, we have given a full exposition of the Ahlswede-Winter Method. Here, we follow a different path due to [53].

## 2.5 Subadditivity of the Matrix Cumulant Generating Function

Let us return to the problem of bounding the matrix moment generating function of an independent sum. Although the multiplicative rule (2.38) for the matrix case is a dead end, the scalar cumulant generating function has a related property that can be extended. For an independent sequence $X_k$ of real random variables, the scalar cumulant generating function is additive:

$$\Xi_{\left(\sum_k X_k\right)}(\theta) = \log\mathbb{E}\exp\left(\sum_k \theta X_k\right) = \sum_k \log\mathbb{E}e^{\theta X_k} = \sum_k \Xi_{X_k}(\theta), \tag{2.40}$$

where the second relation follows from (2.38) when we take logarithms.

The key insight of Tropp's approach is that Corollary 1.4.18 offers a completely way to extend the addition rule (2.40) for the scalar setting to the matrix setting. Indeed, this is a remarkable breakthrough. Much better results have been obtained due to this breakthrough. This justifies the parallel development of Tropp's method with the Ahlswede-Winter method of Sect. 2.2.4.

**Lemma 2.5.1 (Subadditivity of Matrix Cumulant Generating Functions).** *Consider a finite sequence $\{\mathbf{X}_k\}$ of independent, random, Hermitian matrices. Then*

$$\mathbb{E}\operatorname{Tr}\exp\left(\sum_k \theta\mathbf{X}_k\right) \leqslant \operatorname{Tr}\exp\left(\sum_k \log\mathbb{E}e^{\theta\mathbf{X}_k}\right) \textit{for } \theta\in\mathbb{R}. \tag{2.41}$$

*Proof.* It does not harm to assume $\theta = 1$. Let $\mathbb{E}_k$ denote the expectation, conditioned on $\mathbf{X}_1,\ldots,\mathbf{X}_k$. Abbreviate

$$\mathbf{\Xi}_k := \log\left(\mathbb{E}_{k-1}e^{\mathbf{X}_k}\right) = \log\left(\mathbb{E}e^{\mathbf{X}_k}\right),$$

where the equality holds because the sequence $\{\mathbf{X}_k\}$ is independent.

$$\begin{aligned}
\mathbb{E}\operatorname{Tr}\exp\left(\sum_{k=1}^{n}\mathbf{X}_k\right) &= \mathbb{E}_0\cdots\mathbb{E}_{n-1}\operatorname{Tr}\exp\left(\sum_{k=1}^{n-1}\mathbf{X}_k+\mathbf{X}_n\right)\\
&\leqslant \mathbb{E}_0\cdots\mathbb{E}_{n-2}\operatorname{Tr}\exp\left(\sum_{k=1}^{n-1}\mathbf{X}_k+\log\left(\mathbb{E}_{n-1}e^{\mathbf{X}_n}\right)\right)\\
&= \mathbb{E}_0\cdots\mathbb{E}_{n-2}\operatorname{Tr}\exp\left(\sum_{k=1}^{n-2}\mathbf{X}_k+\mathbf{X}_{n-1}+\boldsymbol{\Xi}_n\right)\\
&\leqslant \mathbb{E}_0\cdots\mathbb{E}_{n-3}\operatorname{Tr}\exp\left(\sum_{k=1}^{n-2}\mathbf{X}_k+\boldsymbol{\Xi}_{n-1}+\boldsymbol{\Xi}_n\right)\\
&\cdots\leqslant \operatorname{Tr}\exp\left(\sum_{k=1}^{n}\boldsymbol{\Xi}_k\right).
\end{aligned}$$

The first line follows from the tower property of conditional expectation. At each step, $m = 1, 2, \ldots, n$, we use Corollary 1.4.18 with the fixed matrix $\mathbf{H}$ equal to

$$\mathbf{H}_m = \sum_{k=1}^{m-1}\mathbf{X}_k + \sum_{k=m+1}^{n}\boldsymbol{\Xi}_k.$$

This act is legal because $\mathbf{H}_m$ does not depend on $\mathbf{X}_m$. □

To be in contrast with the additive rule (2.40), we rewrite (2.41) in the form

$$\mathbb{E}\operatorname{Tr}\exp\left(\boldsymbol{\Xi}_{\left(\sum_k \mathbf{X}_k\right)}(\theta)\right) \leqslant \operatorname{Tr}\exp\left(\sum_k \boldsymbol{\Xi}_{\mathbf{X}_k}(\theta)\right) \text{ for } \theta\in\mathbb{R}$$

by using definition (2.37).

## 2.6 Tail Bounds for Independent Sums

This section contains abstract tail bounds for the sums of random matrices.

**Theorem 2.6.1 (Master Tail Bound for Independent Sums—Tropp [53]).** *Consider a finite sequence $\{\mathbf{X}_k\}$ of independent, random, Hermitian matrices. For all $t\in\mathbb{R}$,*

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_k \mathbf{X}_k\right)\geqslant t\right)\leqslant \inf_{\theta>0}\left\{e^{-\theta t}\cdot\operatorname{Tr}\exp\left(\sum_k \log\mathbb{E}e^{\theta\mathbf{X}_k}\right)\right\}. \quad (2.42)$$

*Proof.* Substitute the subadditivity rule for matrix cumulant generating functions, Eq.2.41, into the Lapalace transform bound, Proposition 2.3.1. □

Now we are in a position to apply the very general inequality of (2.42) to some specific situations. The first corollary adapts Theorem 2.6.1 to the case that arises most often in practice.

**Corollary 2.6.2 (Tropp [53]).** *Consider a finite sequence* $\{\mathbf{X}_k\}$ *of independent, random, Hermitian matrices with dimension* $d$*. Assume that there is a function* $g : (0,\infty) \to [0,\infty]$ *and a sequence of* $\{\mathbf{A}_k\}$ *of fixed Hermitian matrices that satisfy the relations*

$$\mathbb{E}e^{\theta\mathbf{X}_k} \leqslant \mathbb{E}e^{g(\theta)\cdot\mathbf{A}_k} \ \textit{for} \ \ \theta > 0. \tag{2.43}$$

*Define the scale parameter*

$$\rho := \lambda_{max}\left(\sum_k \mathbf{A}_k\right).$$

*Then, For all* $t \in \mathbb{R}$,

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_k \mathbf{X}_k\right) \geqslant t\right) \leqslant d \cdot \inf_{\theta>0}\left\{e^{-\theta t + g(\theta)\cdot\rho}\right\}. \tag{2.44}$$

*Proof.* The hypothesis (2.44) implies that

$$\log\mathbb{E}e^{\theta\mathbf{X}_k} \leqslant g(\theta)\cdot\mathbf{A}_k \ \text{ for } \ \theta > 0. \tag{2.45}$$

because of the property (1.73) that the matrix logarithm is operator monotone. Recall the fact (1.70) that the trace exponential is monotone with respect to the semidefinite order. As a result, we can introduce each relation from the sequence (2.45) into the master inequality (2.42). For $\theta > 0$, it follows that

$$\begin{aligned}
\mathbb{P}\left(\lambda_{\max}\left(\sum_k \mathbf{X}_k\right) \geqslant t\right) &\leqslant e^{-\theta t}\cdot \operatorname{Tr}\exp\left(g(\theta)\cdot\sum_k \mathbf{A}_k\right)\\
&\leqslant e^{-\theta t}\cdot d\cdot\lambda_{\max}\left[\exp\left(g(\theta)\cdot\sum_k \mathbf{A}_k\right)\right]\\
&= e^{-\theta t}\cdot d\cdot\exp\left(g(\theta)\cdot\lambda_{\max}\left(\sum_k \mathbf{A}_k\right)\right).
\end{aligned}$$

The second inequality holds because the trace of a positive definite matrix, such as the exponential, is bounded by the dimension $d$ times the maximum eigenvalue. The last line depends on the spectral mapping Theorem 1.4.4 and the fact that the function $g$ is nonnegative. Identify the quantity $\rho$, and take the infimum over positive $\theta$ to reach the conclusion (2.44). □

Let us state another consequence of Theorem 2.6.1. This bound is sometimes more convenient than Corollary 2.6.2, since it combines the matrix generating functions of the random matrices together under a single logarithm.

**Corollary 2.6.3.** *Consider a sequence* $\{\mathbf{X}_k, k = 1, 2, \ldots, n\}$ *of independent, random, Hermitian matrices with dimension* $d$*. For all* $t \in \mathbb{R}$*,*

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_{k=1}^{n}\mathbf{X}_k\right) \geqslant t\right) \leqslant d \cdot \inf_{\theta>0} \exp\left\{-\theta t + n \cdot \log \lambda_{max}\left(\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}e^{\theta \mathbf{X}_k}\right)\right\}. \tag{2.46}$$

*Proof.* Recall the fact (1.74) that the matrix logarithm is operator concave. For $\theta > 0$, it follows that

$$\sum_{k=1}^{n}\log \mathbb{E}e^{\theta \mathbf{X}_k} = n \cdot \frac{1}{n}\sum_{k=1}^{n}\log \mathbb{E}e^{\theta \mathbf{X}_k} \leqslant n \cdot \log\left(\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}e^{\theta \mathbf{X}_k}\right).$$

The property (1.70) that the trace exponential is monotone allows us to introduce the latter relation into the master inequality (2.42) to obtain

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_{k=1}^{n}\mathbf{X}_k\right) \geqslant t\right) \leqslant e^{-\theta t} \cdot \operatorname{Tr}\exp\left(n \cdot \log\left(\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}e^{\theta \mathbf{X}_k}\right)\right).$$

To bound the proof, we bound the trace by $d$ times the maximum eigenvalue, and we invoke the spectral mapping Theorem (twice) 1.4.4 to draw the maximum eigenvalue map inside the logarithm. Take the infimum over positive $\theta$ to reach (2.46).

□

We can study the minimum eigenvalue of a sum of random Hermitian matrices because

$$\lambda_{\min}(\mathbf{X}) = -\lambda_{\max}(-\mathbf{X}).$$

As a result,

$$\mathbb{P}\left(\lambda_{\min}\left(\sum_{k=1}^{n}\mathbf{X}_k\right) \leqslant t\right) = \mathbb{P}\left(\lambda_{\max}\left(\sum_{k=1}^{n}-\mathbf{X}_k\right) \geqslant -t\right).$$

We can also analyze the maximum singular value of a sum of random *rectangular* matrices by applying the results to the Hermitian dilation (1.81). For a finite sequence $\{\mathbf{Z}_k\}$ of independent, random, rectangular matrices, we have

$$\mathbb{P}\left(\left\|\sum_{k}\mathbf{X}_k\right\| \geqslant t\right) = \mathbb{P}\left(\lambda_{\max}\left(\sum_{k}\varphi(\mathbf{Z}_k)\right) \geqslant t\right)$$

on account of (1.83) and the property that the dilation is real-linear. $\varphi$ means dilation. This device allows us to extend most of the tail bounds developed in this book to rectangular matrices.

### 2.6.1 Comparison Between Tropp's Method and Ahlswede–Winter Method

Ahlswede and Winter uses a different approach to bound the matrix moment generating function, which uses the multiplication bound (2.39) for the trace exponential of a sum of two independent, random, Hermitian matrices.

Consider a sequence $\{\mathbf{X}_k, k = 1, 2, \ldots, n\}$ of independent, random, Hermitian matrices with dimension $d$, and let $\mathbf{Y} = \sum_k \mathbf{X}_k$. The trace inequality (2.39) implies that

$$\begin{aligned}
\operatorname{Tr}\mathbf{M}_{\mathbf{Y}}(\theta) &\leqslant \mathbb{E}\operatorname{Tr}\left[e^{\sum\limits_{k=1}^{n-1}\theta\mathbf{X}_k} e^{\theta\mathbf{X}_n}\right] = \operatorname{Tr}\mathbb{E}\left[e^{\sum\limits_{k=1}^{n-1}\theta\mathbf{X}_k} e^{\theta\mathbf{X}_n}\right] \\
&= \operatorname{Tr}\left[\left(\mathbb{E}e^{\sum\limits_{k=1}^{n-1}\theta\mathbf{X}_k}\right)\left(\mathbb{E}e^{\theta\mathbf{X}_n}\right)\right] \\
&\leqslant \operatorname{Tr}\left(\mathbb{E}e^{\sum\limits_{k=1}^{n-1}\theta\mathbf{X}_k}\right)\cdot\lambda_{\max}\left(\mathbb{E}e^{\theta\mathbf{X}_n}\right).
\end{aligned}$$

These steps are carefully spelled out in previous sections, for example Sect. 2.2.4.

Iterating this procedure leads to the relation

$$\operatorname{Tr}\mathbf{M}_{\mathbf{Y}}(\theta) \leqslant (\operatorname{Tr}\mathbf{I})\prod_k \lambda_{\max}\left(\mathbb{E}e^{\theta\mathbf{X}_k}\right) = d\cdot\exp\left(\sum_k \lambda_{\max}\left(\log\mathbb{E}e^{\theta\mathbf{X}_k}\right)\right). \tag{2.47}$$

This bound (2.47) is the key to the Ahlswede–Winter method. As a consequence, their approach generally leads to tail bounds that depend on a scale parameter involving "the sum of eigenvalues." In contrast, the Tropp's approach is based on the subadditivity of cumulants, Eq. 2.41, which implies that

$$\operatorname{Tr}\mathbf{M}_{\mathbf{Y}}(\theta) \leqslant d\cdot\exp\left(\lambda_{\max}\left(\sum_k \log\mathbb{E}e^{\theta\mathbf{X}_k}\right)\right). \tag{2.48}$$

(2.48) contains a scale parameter that involves the "eigenvalues of a sum."

## 2.7 Matrix Gaussian Series—Case Study

A matrix Gaussian series stands among the simplest instances of a sum of independent random matrices. We study this fundamental problem to gain insights.

Consider a finite sequence $a_k$ of real numbers and finite sequence $\{\gamma_k\}$ of independent, standard Gaussian variables. We have

$$\mathbb{P}\left(\sum_k \gamma_k a_k \geqslant t\right) \leqslant e^{-t^2/2\sigma^2} \quad \text{where} \quad \sigma^2 := \sum_k a_k^2. \tag{2.49}$$

This result justifies that a Gaussian series with real coefficients satisfies a normal-type tail bound where the variance is controlled by the sum of the sequence coefficients. The relation (2.49) follows easily from the scalar Laplace transform method. See Example 1.2.1 for the derivation of the characteristic function; A Fourier inverse transform of this derived characteristic function will lead to (2.49). So far, our exposition in this section is based on the standard textbook.

The inequality (2.49) can be generalized directly to the noncommutative setting. The matrix Laplace transform method, Proposition 2.3.1, delivers the following result.

**Theorem 2.7.1 (Matrix Gaussian and Rademacher Series—Tropp [53]).** *Consider a finite sequence $\{\mathbf{A}_k\}$ of fixed Hermitian matrices with dimension $d$, and let $\gamma_k$ be a finite sequence of independent standard normal variables. Compute the variance parameter*

$$\sigma^2 := \left\| \sum_k \mathbf{A}_k^2 \right\|. \tag{2.50}$$

*Then, for all $t > 0$,*

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_k \gamma_k \mathbf{A}_k\right) \geqslant t\right) \leqslant d \cdot e^{-t^2/2\sigma^2}. \tag{2.51}$$

*In particular,*

$$\mathbb{P}\left(\left\|\sum_k \gamma_k \mathbf{A}_k\right\| \geqslant t\right) \leqslant 2d \cdot e^{-t^2/2\sigma^2}. \tag{2.52}$$

*The same bounds hold when we replace $\gamma_k$ by a finite sequence of independent Rademacher random variables.*

Observe that the bound (2.51) reduces to the scalar result (2.49) when the dimension $d = 1$. The generalization of (2.50) has been proven by Tropp [53] to be sharp and is also demonstrated that Theorem 2.7.1 cannot be improved without changing its form.

Most of the inequalities in this book have variants that concern the maximum singular value of a sum of *rectangular* random matrices. These extensions follow immediately, as mentioned above, when we apply the Hermitian matrices to the Hermitian dilation of the sums of rectangular matrices. Here is a general version of Theorem 2.7.1.

**Corollary 2.7.2 (Rectangular Matrix Gaussian and Radamacher Series—Tropp [53]).** *Consider a finite sequence $\{\mathbf{B}_k\}$ of fixed matrices with dimension $d_1 \times d_2$, and let $\gamma_k$ be a finite sequence of independent standard normal variables. Compute the variance parameter*

$$\sigma^2 := \max\left\{\left\|\sum_k \mathbf{B}_k\mathbf{B}_k^*\right\|, \left\|\sum_k \mathbf{B}_k^*\mathbf{B}_k\right\|\right\}.$$

*Then, for all $t > 0$,*

$$\mathbb{P}\left(\left\|\sum_k \gamma_k \mathbf{B}_k\right\| \geqslant t\right) \leqslant (d_1 + d_2)\cdot e^{-t^2/2\sigma^2}.$$

*The same bounds hold when we replace $\gamma_k$ by a finite sequence of independent Rademacher random variables.*

To prove Theorem 2.7.1 and Corollary 2.7.2, we need a lemma first.

**Lemma 2.7.3 (Rademacher and Gaussian moment generating functions).** *Suppose that $\mathbf{A}$ is an Hermitian matrix. Let $\varepsilon$ be a Rademacher random variable, and let $\gamma$ be a standard normal random variable. Then*

$$\mathbb{E}e^{\varepsilon\theta\mathbf{A}} \leqslant e^{\theta^2\mathbf{A}^2/2} \;\; \textit{and} \;\; \mathbb{E}e^{\gamma\theta\mathbf{A}} = e^{\theta^2\mathbf{A}^2/2} \;\; \textit{for} \;\; \theta \in \mathbb{R}.$$

*Proof.* Absorbing $\theta$ into $\mathbf{A}$, we may assume $\theta = 1$ in each case. By direct calculation,

$$\mathbb{E}e^{\varepsilon\mathbf{A}} = \cosh(\mathbf{A}) \leqslant e^{\mathbf{A}^2/2},$$

where the second relation is (1.69).

For the Gaussian case, recall that the moments of a standard normal variable satisfy

$$\mathbb{E}\gamma^{2k+1} = 0 \;\; \text{and} \;\; \mathbb{E}\gamma^{2k} = \frac{(2k)!}{k!2^k} \;\; \text{for} \;\; k = 0, 1, 2, \ldots.$$

Therefore,

$$\mathbb{E}e^{\gamma\mathbf{A}} = \mathbf{I} + \sum_{k=1}^{\infty} \frac{\mathbb{E}\left(\gamma^{2k}\right)\mathbf{A}^{2k}}{(2k)!} = \mathbf{I} + \sum_{k=1}^{\infty} \frac{\left(\mathbf{A}^2/2\right)^k}{k!} = e^{\mathbf{A}^2/2}.$$

The first relation holds since the odd terms in the series vanish. With this lemma, the tail bounds for Hermitian matrix Gaussian and Rademacher series follow easily. □

*Proof of Theorem 2.7.1*. Let $\{\xi_k\}$ be a finite sequence of independent standard normal variables or independent Rademacher variables. Invoke Lemma 2.7.3 to obtain

$$\mathbb{E}e^{\xi_k\theta\mathbf{A}} \leqslant e^{g(\theta)\cdot\mathbf{A}_k^2} \quad \text{where} \quad g\left(\theta\right) := \theta^2/2 \;\; \text{for} \;\; \theta > 0.$$

Recall that

$$\sigma^2 = \left\|\sum_k \mathbf{A}_k^2\right\| = \lambda_{\max}\left(\sum_k \mathbf{A}_k^2\right).$$

Corollary 2.6.2 gives

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_k \xi_k\mathbf{A}_k\right) \geqslant t\right) \leqslant d\cdot \inf_{\theta>0}\left\{e^{-\theta t + g(\theta)\cdot\sigma^2}\right\} = d\cdot e^{-t^2/2\sigma^2}. \quad (2.53)$$

For the record, the infimum is attained when $\theta = t/\sigma^2$.

To obtain the norm bound (2.52), recall that

$$\|\mathbf{Y}\| = \max\left\{\lambda_{\max}\left(\mathbf{Y}\right), -\lambda_{\min}\left(\mathbf{Y}\right)\right\}.$$

Since standard Gaussian and Rademacher variables are symmetric, the inequality (2.53) implies that

$$\mathbb{P}\left(-\lambda_{\min}\left(\sum_k \xi_k\mathbf{A}_k\right) \geqslant t\right) = \mathbb{P}\left(\lambda_{\max}\left(\sum_k \left(-\xi_k\right)\mathbf{A}_k\right) \geqslant t\right) \leqslant d\cdot e^{-t^2/2\sigma^2}.$$

Apply the union bound to the estimates for $\lambda_{\max}$ and $-\lambda_{\min}$ to complete the proof. We use the Hermitian dilation of the series. □

*Proof of Corollary 2.7.2*. Let $\{\xi_k\}$ be a finite sequence of independent standard normal variables or independent Rademacher variables. Consider the sequence $\{\xi_k\varphi\left(\mathbf{B}_k\right)\}$ of random Hermitian matrices with dimension $d_1 + d_2$. The spectral identity (1.83) ensures that

$$\left\| \sum_k \xi_k \mathbf{B}_k \right\| = \lambda_{\max} \left( \varphi \left( \sum_k \xi_k \mathbf{B}_k \right) \right) = \lambda_{\max} \left( \sum_k \xi_k \varphi \left( \mathbf{B}_k \right) \right).$$

Theorem 2.7.1 is used. Simply observe that the matrix variance parameter (2.50) satisfies the relation

$$\sigma^2 = \left\| \sum_k \varphi(\mathbf{B}_k)^2 \right\| = \left\| \begin{bmatrix} \sum_k \mathbf{B}_k \mathbf{B}_k^* & \mathbf{0} \\ \mathbf{0} & \sum_k \mathbf{B}_k^* \mathbf{B}_k \end{bmatrix} \right\|$$
$$= \max \left\{ \left\| \sum_k \mathbf{B}_k \mathbf{B}_k^* \right\|, \left\| \sum_k \mathbf{B}_k^* \mathbf{B}_k \right\| \right\}.$$

on account of the identity (1.82). □

## 2.8 Application: A Gaussian Matrix with Nonuniform Variances

Fix a $d_1 \times d_2$ matrix $\mathbf{B}$ and draw a random $d_1 \times d_2$ matrix $\mathbf{\Gamma}$ whose entries are independent, standard normal variables. Let $\odot$ denote the componentwise (i.e., Schur or Hadamard) product of matrices. Construct the random matrix $\mathbf{B} \odot \mathbf{\Gamma}$, and observe that its $(i,j)$ component is a Gaussian variable with mean zero and variance $|b_{ij}|^2$. We claim that

$$\mathbb{P}\left\{ \|\Gamma \odot \mathbf{B}\| \geqslant t \right\} \leqslant (d_1 + d_2) \cdot e^{-t^2/2\sigma^2}. \tag{2.54}$$

The symbols $\mathbf{b}_{i:}$ and $\mathbf{b}_{:j}$ represent the $i$th row and $j$th column of the matrix $\mathbf{B}$. An immediate sequence of (2.54) is that the median of the norm satisfies

$$\mathbb{M}\left( \|\Gamma \odot \mathbf{B}\| \right) \leqslant \sigma \sqrt{2 \log \left( 2 \left( d_1 + d_2 \right) \right)}. \tag{2.55}$$

These are nonuniform Gaussian matrices where the estimate (2.55) for the median has the correct order. We compare [106, Theorem 1] and [107, Theorem 3.1] although the results are not fully comparable. See Sect. 9.2.2 for extended work.

To establish (2.54), we first decompose the matrix of interest as a Gaussian series:

$$\mathbf{\Gamma} \odot \mathbf{B} = \sum_{ij} \gamma_{ij} \cdot b_{ij} \cdot \mathbf{C}_{ij}.$$

Now, let us determine the variance parameter $\sigma^2$. Note that

$$\sum_{ij} (b_{ij}\mathbf{C}_{ij})(b_{ij}\mathbf{C}_{ij})^* = \sum_i \left( \sum_j |b_{ij}|^2 \right) \mathbf{C}_{ii} = \mathrm{diag}\left( \|\mathbf{b}_{1:}\|^2, \ldots, \|\mathbf{b}_{d_1:}\|^2 \right).$$

Similarly,

$$\sum_{ij} (b_{ij}\mathbf{C}_{ij})^* (b_{ij}\mathbf{C}_{ij}) = \sum_j \left( \sum_i |b_{ij}|^2 \right) \mathbf{C}_{jj} = \mathrm{diag}\left( \|\mathbf{b}_{:1}\|^2, \ldots, \|\mathbf{b}_{:d_2}\|^2 \right).$$

Thus,

$$\begin{aligned}\sigma^2 &= \max\left\{ \left\| \mathrm{diag}\left( \|\mathbf{b}_{1:}\|^2, \ldots, \|\mathbf{b}_{d_1:}\|^2 \right) \right\|, \left\| \mathrm{diag}\left( \|\mathbf{b}_{:1}\|^2, \ldots, \|\mathbf{b}_{:d_2}\|^2 \right) \right\| \right\} \\ &= \max\left\{ \max_i \|\mathbf{b}_{i:}\|^2, \max_j \|\mathbf{b}_{:j}\|^2 \right\}.\end{aligned}$$

An application of Corollary 2.7.2 gives the tail bound of (2.54).

## 2.9 Controlling the Expectation

The Hermitian Gaussian series

$$\mathbf{Y} = \sum_k \gamma_k \mathbf{A}_k \tag{2.56}$$

is used for many practical applications later in this book since it allows each sensor to be represented by the $k$th matrix.

*Example 2.9.1 (NC-OFDM Radar and Communications).* A subcarrier (or tone) has a frequency $f_k, k = 1, \ldots, N$. Typically, $N = 64$ or $N = 128$. A radio sinusoid $e^{j2\pi f_k t}$ is transmitted by the transmitter (cell phone tower or radar). This radio signal passes through the radio environment and "senses" the environment. Each sensor collects some length of data over the sensing time. The data vector $\mathbf{y}_k$ of length $10^6$ is stored and processed for only one sensor. In other words, we receive typically $N = 128$ copies of measurements for using one sensor. Of course, we can use more sensors, say $M = 100$.

We can extract the data structure using a covariance matrix that is to be directly estimated from the data. For example, a sample covariance matrix can be used. We call the estimated covariance matrix $\hat{\mathbf{R}}_k, k = 1, 2, .., N$. We may desire to know the impact of $N$ subcarriers on the sensing performance. Equation (2.56) is a natural model for this problem at hand. If we want to investigate the impact of $M = 100$ sensors on the sensing performance (via collaboration from a wireless network), we need a data fusion algorithm. Intuitively, we can simply consider the sum of these extracted covariance matrices (random matrices). So we have a total of $n = MN = 100 \times 128 = 12{,}800$ random matrices at our disposal. Formally, we have

$$\mathbf{Y} = \sum_k \gamma_k \mathbf{A}_k = \sum_{k=1}^{n=128{,}000} \gamma_k \hat{\mathbf{R}}_k.$$

Here, we are interested in the nonasymptotic view in statistics [108]: when the number of observations $n$ is large, we fit large complex sets of data that one needs to deal with huge collections of models at different scales. Throughout the book, we promote this nonasymptotic view by solving practical problems in wireless sensing and communications. This is a problem with "Big Data". In this novel view, one takes the number of observations as it is and try to evaluate the effect of all the influential parameters. Here this parameter is $n$, the total number of measurements. Within one second, we have a total of $10^6 \times 128 \times 100 \approx 10^{10}$ points of data at our disposal. We need models at different scales to represent the data. $\square$

A remarkable feature of Theorem 2.7.1 is that it always allows us to obtain reasonably accurate estimates for the expected norm of this Hermitian Gaussian series

$$\mathbf{Y} = \sum_k \gamma_k \mathbf{A}_k. \tag{2.57}$$

To establish this point, we first derive the upper and lower bounds for the second moment of $||\mathbf{Y}||$. Note $||\mathbf{Y}||$ is a scalar random variable. Using Theorem 2.7.1 gives

$$\begin{aligned}\mathbb{E}\left(\|\mathbf{Y}\|^2\right) &= \int_0^\infty \mathbb{P}\left(\|\mathbf{Y}\| > \sqrt{t}\right)dt \\ &= 2\sigma^2 \log(2d) + 2d\int_{2\sigma^2\log(2d)}^\infty e^{-t/2\sigma^2}dt = 2\sigma^2\log(2ed)\,.\end{aligned}$$

Jensen's inequality furnishes the lower estimate:

$$\mathbb{E}\left(\|\mathbf{Y}\|^2\right) = \mathbb{E}\left(\left\|\mathbf{Y}^2\right\|\right) \geqslant \left\|\mathbb{E}\mathbf{Y}^2\right\| = \left\|\sum_k \mathbf{A}_k^2\right\| = \sigma^2.$$

The (homogeneous) first and second moments of the norm of a Gaussian series are equivalent up to a universal constant [109, Corollary 3.2], so we have

$$c\sigma \leqslant \mathbb{E}\left(\|\mathbf{Y}\|\right) \leqslant \sigma\sqrt{2\log(2ed)}. \tag{2.58}$$

According to (2.58), the matrix variance parameter $\sigma^2$ controls the expected norm $\mathbb{E}\left(\|\mathbf{Y}\|\right)$ up to a factor that depends very weakly on the dimension $d$. A similar remark goes to the median value $\mathbb{M}\left(\|\mathbf{Y}\|\right)$.

In (2.58), the dimensional dependence is a new feature of probability inequalities in the matrix setting. We cannot remove the factor $d$ from the bound in Theorem 2.7.1.

## 2.10 Sums of Random Positive Semidefinite Matrices

The classical Chernoff bounds concern the sum of independent, nonnegative, and *uniformly bounded* random variables. In contrast, matrix Chernoff bounds deal with a sum of independent, positive semidefinite, random matrices whose maximum eigenvalues are subject to a uniform bound. For example, the sample covariance matrices satisfy the conditions of independent, positive semidefinite, random matrices. This connection plays a fundamental role when we deal with cognitive sensing in the network setting consisting of a number of sensors. Roughly, each sensor can be modeled by a sample covariance matrix.

The first result parallels with the strongest version of the scalar Chernoff inequality for the proportion of successes in a sequence of independent, (but not identical) Bernoulli trials [7, Excercise 7].

**Theorem 2.10.1 (Matrix Chernoff I—Tropp [53]).** *Consider a sequence* $\{\mathbf{X}_k : k = 1, \ldots, n\}$ *of independent, random, Hermitian matrices that satisfy*

$$\mathbf{X}_k \geqslant 0 \quad \text{and} \quad \lambda_{\max}\left(\mathbf{X}_k\right) \leqslant 1 \quad \text{almost surely.}$$

*Compute the minimum and maximum eigenvalues of the average expectation,*

$$\bar{\mu}_{\min} := \lambda_{\min}\left(\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\mathbf{X}_k\right) \quad \text{and} \quad \bar{\mu}_{\max} := \lambda_{\max}\left(\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\mathbf{X}_k\right).$$

*Then*

$$\begin{aligned}
\mathbb{P}\left\{\lambda_{\min}\left(\tfrac{1}{n}\sum_{k=1}^{n}\mathbf{X}_k\right) \leqslant \alpha\right\} &\leqslant d \cdot e^{-n \cdot D(\alpha||\bar{\mu}_{\min})} \quad \text{for} \quad 0 \leqslant \alpha \leqslant \bar{\mu}_{\min}, \\
\mathbb{P}\left\{\lambda_{\max}\left(\tfrac{1}{n}\sum_{k=1}^{n}\mathbf{X}_k\right) \geqslant \alpha\right\} &\leqslant d \cdot e^{-n \cdot D(\alpha||\bar{\mu}_{\max})} \quad \text{for} \quad 0 \leqslant \alpha \leqslant \bar{\mu}_{\max}.
\end{aligned}$$

*the binary information divergence*

$$D\left(a||u\right) = a\left(\log\left(a\right) - \log\left(u\right)\right) + \left(1-a\right)\left(\log\left(1-a\right) - \log\left(1-u\right)\right)$$

*for* $a, u \in [0, 1]$.

Tropp [53] found the following weaker version of Theorem 2.10.1 produces excellent results but is simpler to apply. This corollary corresponds with the usual statement of the scalar Chernoff inequalities for sums of nonnegative random variables; see [7, Excercise 8] [110, Sect. 4.1]. Theorem 2.10.1 is a considerable strengthening of the version of Ahlswede-Winter [36, Theorem 19], in which case their result requires the assumption that the summands are identically distributed.

**Corollary 2.10.2 (Matrix Chernoff II—Tropp [53]).** *Consider a sequence* $\{\mathbf{X}_k : k = 1, \ldots, n\}$ *of independent, random, Hermitian matrices that satisfy*

$$\mathbf{X}_k \geqslant 0 \quad \textit{and} \quad \lambda_{\max}\left(\mathbf{X}_k\right) \leqslant R \quad \textit{almost surely}.$$

*Compute the minimum and maximum eigenvalues of the average expectation,*

$$\mu_{\min} := \lambda_{\min}\left(\sum_{k=1}^{n} \mathbb{E}\mathbf{X}_k\right) \ \ \textit{and} \ \ \mu_{\max} := \lambda_{\max}\left(\sum_{k=1}^{n} \mathbb{E}\mathbf{X}_k\right).$$

*Then*

$$\begin{gathered}\mathbb{P}\left\{\lambda_{\min}\left(\sum_{k=1}^{n}\mathbf{X}_k\right) \leqslant (1-\delta)\,\mu_{\min}\right\} \leqslant d \cdot \left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\mu_{\min}/R} \textit{for } \delta \in [0,1], \\ \mathbb{P}\left\{\lambda_{\max}\left(\sum_{k=1}^{n}\mathbf{X}_k\right) \geqslant (1+\delta)\,\mu_{\max}\right\} \leqslant d \cdot \left[\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right]^{\mu_{\max}/R} \textit{for } \delta \geqslant 0.\end{gathered}$$

The following standard simplification of Corollary 2.10.2 is useful:

$$\begin{gathered}\mathbb{P}\left\{\lambda_{\min}\left(\sum_{k=1}^{n}\mathbf{X}_k\right) \leqslant t\mu_{\min}\right\} \leqslant d \cdot e^{-(1-t)^2\mu_{\min}/2R} \ \text{ for } \ \mathrm{t} \in [0,1], \\ \mathbb{P}\left\{\lambda_{\max}\left(\sum_{k=1}^{n}\mathbf{X}_k\right) \geqslant t\mu_{\max}\right\} \leqslant d \cdot \left[\frac{e}{t}\right]^{t\mu_{\max}/R} \ \text{ for } \ \mathrm{t} \geqslant e.\end{gathered}$$

The minimum eigenvalues has norm-type behavior while the maximum eigenvalues exhibits Poisson-type decay.

Before giving the proofs, we consider applications.

*Example 2.10.3 (Rectangular Random Matrix).* Matrix Chernoff inequalities are very effective for studying random matrices with independent columns. Consider a rectangular random matrix

$$\mathbf{Z} = \left[\, \mathbf{z}_1 \ \mathbf{z}_2 \ \cdots \ \mathbf{z}_n \,\right]$$

where $\{\mathbf{z}_k\}$ is a family of independent random vector in $\mathbb{C}^m$. The sample covariance matrix is defined as

$$\hat{\mathbf{R}} = \frac{1}{n}\mathbf{Z}\mathbf{Z}^* = \frac{1}{n}\sum_{k=1}^{n}\mathbf{z}_k\mathbf{z}_k^*,$$

which is an estimate of the true covariance matrix $\mathbf{R}$. One is interested in the error $\left\| \hat{\mathbf{R}} - \mathbf{R} \right\|$ as a function of the number of sample vectors, $n$. The norm of $\mathbf{Z}$ satisfies

$$\|\mathbf{Z}\|^2 = \lambda_{\max}\left(\mathbf{Z}\mathbf{Z}^*\right) = \lambda_{\max}\left(\sum_{k=1}^{n} \mathbf{z}_k \mathbf{z}_k^*\right). \tag{2.59}$$

Similarly, the minimum singular value $s_m$ of the matrix satisfies

$$s_m(\mathbf{Z})^2 = \lambda_{\min}\left(\mathbf{Z}\mathbf{Z}^*\right) = \lambda_{\min}\left(\sum_{k=1}^{n} \mathbf{z}_k \mathbf{z}_k^*\right).$$

In each case, the summands are stochastically independent and positive semidefinite (rank 1) matrices, so the matrix Chernoff bounds apply. □

Corollary 2.10.2 gives accurate estimates for the expectation of the maximum eigenvalue:

$$\mu_{\max} \leqslant \mathbb{E}\lambda_{\max}\left(\sum_{k=1}^{n} \mathbf{X}_k\right) \leqslant C \cdot \max\left\{\mu_{\max}, R\log d\right\}. \tag{2.60}$$

The lower bound is Jensen's inequality; the upper bound is from a standard calculation. The dimensional dependence vanishes, when the mean $\mu_{\max}$ is sufficiently large in comparison with the upper bound $R$! The a prior knowledge of knowing $R$ accurately in $\lambda_{\max}\left(\mathbf{X}_k\right) \leqslant R$ converts into the tighter bound in (2.60).

*Proof of Theorem 2.10.1.* We start with a semidefinite bound for the matrix moment generating function of a random positive semidefinite contraction.

**Lemma 2.10.4 (Chernoff moment generating function).** *Suppose that* $\mathbf{X}$ *is a random positive semidefinite matrix that satisfies* $\lambda_{\max}\left(\mathbf{X}_k\right) \leqslant 1$. *Then*

$$\mathbb{E}\left(e^{\theta\mathbf{X}}\right) \leqslant \mathbf{I} + \left(e^{\theta} - 1\right)\left(\mathbb{E}\mathbf{X}\right) \quad \textit{for} \quad \theta \in \mathbb{R}.$$

The proof of Lemma 2.10.4 parallels the classical argument; the matrix adaptation is due to Ashlwede and Winter [36], which is followed in the proof of Theorem 2.2.15.

*Proof of Lemma 2.10.4.* Consider the function

$$f(x) = e^{\theta x}.$$

Since $f$ is convex, its graph has below the chord connecting two points. In particular,

$$f(x) \leqslant f(0) + \left[f(1) - f(0)\right] \cdot x \quad \text{for} \quad x \in [0, 1].$$

More explicitly,

$$e^{\theta x} \leqslant 1 + \left(e^{\theta} - 1\right) \cdot x \text{ for } x \in [0,1].$$

The eigenvalues of $\mathbf{X}$ lie in the interval of $[0,1]$, so the transfer rule (1.61) implies that

$$e^{\theta \mathbf{X}} \leqslant \mathbf{I} + \left(e^{\theta} - 1\right)\mathbf{X}.$$

Expectation respects the semidefinite order, so

$$\mathbb{E}e^{\theta \mathbf{X}} \leqslant \mathbf{I} + \left(e^{\theta} - 1\right)(\mathbb{E}\mathbf{X}).$$

This is the advertised result of Lemma 2.10.4. □

*Proof Theorem 2.10.1, Upper Bound.* The Chernoff moment generating function, Lemma 2.10.4, states that

$$\mathbb{E}e^{\theta \mathbf{X}_k} \leqslant \mathbf{I} + g(\theta)(\mathbb{E}\mathbf{X}_k) \text{ where } g(\theta) = \left(e^{\theta} - 1\right) \text{ for } \theta > 0.$$

As a result, Corollary 2.6.3 implies that

$$\begin{aligned}\mathbb{P}\left\{\lambda_{\max}\left(\sum_{k=1}^{n}\mathbf{X}_k\right) \geqslant t\right\} &\leqslant d \cdot \exp\left(-\theta t + n \cdot \log \cdot \lambda_{\max}\left(\frac{1}{n}\sum_{k=1}^{n}\left(\mathbf{I} + g(\theta)(\mathbb{E}\mathbf{X}_k)\right)\right)\right)\\ &= d \cdot \exp\left(-\theta t + n \cdot \log \cdot \lambda_{\max}\left(\mathbf{I} + g(\theta) \cdot \frac{1}{n}\sum_{k=1}^{n}\left((\mathbb{E}\mathbf{X}_k)\right)\right)\right)\\ &= d \cdot \exp\left(-\theta t + n \cdot \log \cdot \left(1 + g(\theta) \cdot \bar{\mu}_{\max}\right)\right).\end{aligned} \tag{2.61}$$

The third line follows from the spectral mapping Theorem 1.4.13 and the definition of $\bar{\mu}_{\max}$. Make the change of variable $t \mapsto n\alpha$. The right-hand side is smallest when

$$\theta = \log\left(\alpha/(1-\alpha)\right) - \log\left(\bar{\mu}_{\max}/(1-\bar{\mu}_{\max})\right).$$

After substituting these quantiles into (2.61), we obtain the information divergence upper bound. □

*Proof Corollary 2.10.2, Upper Bound.* Assume that the summands satisfy the uniform eigenvalue bound with $R = 1$; the general result follows by re-scaling. The shortest route to the weaker Chernoff bound starts at (2.61). The numerical inequality $\log(1+x) \leq x$, valid for $x > -1$, implies that

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum_{k=1}^{n}\mathbf{X}_k\right) \geqslant t\right\} \leqslant d \cdot \exp\left(-\theta t + g(\theta) \cdot n\bar{\mu}_{\max}\right) = d \cdot \exp\left(-\theta t + g(\theta) \cdot \mu_{\max}\right).$$

Make the change of variable $t \mapsto (1+\delta)\mu_{\max}$, and select the parameter $\theta = \log(1+\delta)$. Simplify the resulting tail bound to complete the proof. □

The lower bounds follow from a closely related arguments.

*Proof Theorem 2.10.1, Lower Bound.* Our starting point is Corollary 2.6.3, considering the sequence $\{-\mathbf{X}_k\}$. In this case, the Chernoff moment generating function, Lemma 2.10.4, states that

$$\mathbb{E}e^{\theta(-\mathbf{X}_k)} = \mathbb{E}e^{(-\theta)\mathbf{X}_k} \leqslant \mathbf{I} - g(\theta)\cdot(\mathbb{E}\mathbf{X}_k) \quad \text{where} \quad g(\theta) = 1 - e^{-\theta} \quad \text{for} \quad \theta > 0.$$

Since $\lambda_{\min}(-\mathbf{A}) = -\lambda_{\max}(\mathbf{A})$, we can again use Corollary 2.6.3 as follows.

$$\begin{aligned}
\mathbb{P}\left\{\lambda_{\min}\left(\sum_{k=1}^{n}\mathbf{X}_k\right) \leqslant t\right\} &= \mathbb{P}\left\{\lambda_{\max}\left(\sum_{k=1}^{n}(-\mathbf{X}_k)\right) \geqslant -t\right\} \\
&\leqslant d\cdot\exp\left(\theta t + n\cdot\log\lambda_{\max}\left(\frac{1}{n}\sum_{k=1}^{n}(\mathbf{I} - g(\theta)\cdot\mathbb{E}\mathbf{X}_k)\right)\right) \\
&= d\cdot\exp\left(\theta t + n\cdot\log\left(1 - g(\theta)\cdot\lambda_{\min}\left(\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\mathbf{X}_k\right)\right)\right) \\
&= d\cdot\exp\left(\theta t + n\cdot\log\left(1 - g(\theta)\cdot\bar{\mu}_{\min}\right)\right).
\end{aligned} \tag{2.62}$$

Make the substitution $t \mapsto n\alpha$. The right-hand side is minimum when

$$\theta = \log\left(\bar{\mu}_{\min}/\left(1 - \bar{\mu}_{\min}\right)\right) - \log\left(\alpha/\left(1-\alpha\right)\right).$$

These steps result in the information divergence lower bound. □

*Proof Corollary 2.10.2, Lower Bound.* As before, assume that the unform bound $R = 1$. We obtain the weaker lower bound as a consequence of (2.62). The numerical inequality $\log(1+x) \le x$, is valid for $x > -1$, so we have

$$\mathbb{P}\left\{\lambda_{\min}\left(\sum_{k=1}^{n}\mathbf{X}_k\right) \leqslant t\right\} \leqslant d\cdot\exp\left(\theta t - g(\theta)\cdot n\bar{\mu}_{\min}\right) = d\cdot\exp\left(\theta t - g(\theta)\cdot\mu_{\min}\right).$$

Make the substitution $t \mapsto (1-\delta)\,\mu_{\min}$, and select the parameter $\theta = -\log(1-\delta)$ to complete the proof. □

## 2.11 Matrix Bennett and Bernstein Inequalities

In the scalar setting, Bennett and Bernstein inequalities deal with a sum of independent, zero-mean random variables that are either bounded or subexponential. In the matrix setting, the analogous results concern a sum of *zero-mean random matrices*. Recall that the classical Chernoff bounds concern the sum of independent, nonnegative, and uniformly bounded random variables while, matrix Chernoff bounds deal with a sum of independent, *positive semidefinite*, random matrices whose maximum eigenvalues are subject to a uniform bound. Let us consider a motivating example first.

*Example 2.11.1 (Signal plus Noise Model).* For example, the sample covariance matrices of Gaussian noise, $\hat{\mathbf{R}}_{ww}$, satisfy the conditions of independent, zero-mean, random matrices. Formally

$$\hat{\mathbf{R}}_{yy} = \hat{\mathbf{R}}_{xx} + \hat{\mathbf{R}}_{ww},$$

$\hat{\mathbf{R}}_{yy}$ represent the sample covariance matrix of the received signal plus noise and $\hat{\mathbf{R}}_{xx}$ of the signal. Apparently, $\hat{\mathbf{R}}_{ww}$, is a zero-mean random matrix. All these matrices are independent, nonnegative, random matrices. □

Our first result considers the case where the maximum eigenvalue of each summand satisfies a uniform bound. Recall from Example 2.10.3 that the norm of a rectangular random matrix $\mathbf{Z}$ satisfies

$$\|\mathbf{Z}\|^2 = \lambda_{\max}\left(\mathbf{Z}\mathbf{Z}^*\right) = \lambda_{\max}\left(\sum_{k=1}^{n} \mathbf{z}_k \mathbf{z}_k^*\right). \tag{2.63}$$

Physically, we can call the norm as the power.

*Example 2.11.2 (Transmitters with bounded power).* Consider a practical application. Assume each transmitter is modeled as the random matrix $\{\mathbf{Z}_k\}, k = 1, \ldots, n$. We have the a prior knowledge that its transmission is bounded in some manner. A model is to consider

$$\mathbb{E}\mathbf{Z}_k = \mathbf{M}; \;\; \lambda_{\max}\left(\mathbf{Z}_k\right) \leqslant R_1, \;\; k = 1, 2, \ldots, n.$$

After the multi-path channel propagation with fading, the constraints become

$$\mathbb{E}\mathbf{X}_k = \mathbf{N}; \;\; \lambda_{\max}\left(\mathbf{X}_k\right) \leqslant R_2, \;\; k = 1, 2, \ldots, n.$$

Without loss of generality, we can always considered the centered matrix-valued random variable

$$\mathbb{E}\mathbf{X}_k = \mathbf{0}; \;\; \lambda_{\max}\left(\mathbf{X}_k\right) \leqslant R, \;\; k = 1, 2, \ldots, n.$$

When a number of transmitters, say $n$, are emitting at the same time, the total received signal is described by

$$\mathbf{Y} = \mathbf{X}_1 + \cdots, \mathbf{X}_n = \sum_{k=1}^{n} \mathbf{X}_k.$$

□

**Theorem 2.11.3 (Matrix Bernstein:Bounded Case—Theorem 6.1 of Tropp [53]).** *Consider a finite sequence* $\{\mathbf{X}_k\}$ *of independent, random, Hermitian matrices with dimension d. Assume that*

$$\mathbb{E}\mathbf{X}_k = \mathbf{0}; \lambda_{\max}(\mathbf{X}_k) \leqslant R, \quad \textit{almost surely}.$$

*Compute the norm of the total variance,*

$$\sigma^2 := \left\| \sum_{k=1}^{n} \mathbb{E}\left(\mathbf{X}_k^2\right) \right\|.$$

*Then the following chain of inequalities holds for all* $t \geq 0$.

$$\begin{aligned} \mathbb{P}\left\{ \lambda_{\max}\left( \sum_{k=1}^{n} \mathbf{X}_k \right) \geqslant t \right\} &\leqslant d \cdot \exp\left( -\tfrac{\sigma^2}{R^2} \cdot h\left( \tfrac{Rt}{\sigma^2} \right) \right) & (i) \\ &\leqslant d \cdot \exp\left( -\tfrac{t^2/2}{\sigma^2 + Rt/3} \right) & (ii) \\ &\leqslant \begin{cases} d \cdot \exp\left(-3t^2/8\sigma^2\right) & \textit{for } t \leqslant \sigma^2/R; \\ d \cdot \exp\left(-3t/8R\right) & \textit{for } t \geqslant \sigma^2/R. \end{cases} & (iii) \end{aligned} \tag{2.64}$$

*The function* $h(x) := (1+x)\log(1+x) - x$ *for* $x \geqslant 0$.

**Theorem 2.11.4 (Matrix Bernstein: Subexponential Case—Theorem 6.2 of Tropp [53]).** *Consider a finite sequence* $\{\mathbf{X}_k\}$ *of independent, random, Hermitian matrices with dimension d. Assume that*

$$\mathbb{E}\mathbf{X}_k = \mathbf{0}; \mathbb{E}\left(\mathbf{X}_k^p\right) \leqslant \frac{p!}{2!} \cdot R^{p-2}\mathbf{A}_k^2, \textit{for } p = 2,3,4,\ldots.$$

*Compute the variance parameter*

$$\sigma^2 := \left\| \sum_{k=1}^{n} \mathbf{A}_k^2 \right\|.$$

*Then the following chain of inequalities holds for all* $t \geq 0$.

$$\begin{aligned} \mathbb{P}\left\{ \lambda_{\max}\left( \sum_{k=1}^{n} \mathbf{X}_k \right) \geq t \right\} &\leqslant d \cdot \exp\left( -\tfrac{t^2/2}{\sigma^2 + Rt} \right) \\ &\leqslant \begin{cases} d \cdot \exp\left(-t^2/4\sigma^2\right) & \textit{for } t \leqslant \sigma^2/R; \\ d \cdot \exp\left(-t/4R\right) & \textit{for } t \geqslant \sigma^2/R. \end{cases} \end{aligned}$$

## 2.12 Minimax Matrix Laplace Method

This section, taking material from [43, 49, 53, 111], combines the matrix Laplace transform method of Sect. 2.3 with the Courant-Fischer characterization of eigenvalues (Theorem 1.4.22) to obtain nontrivial bounds on the interior eigenvalues of a sum of random Hermitian matrices. We will use this approach for estimates of the covariance matrix.

## 2.13 Tail Bounds for All Eigenvalues of a Sum of Random Matrices

In this section, closely following Tropp [111], we develop a generic bound on the tail probabilities of eigenvalues of sums of independent, random, Hermitian matrices. We establish this bound by supplementing the matrix Laplace transform methodology of Tropp [53], that is treated before in Sect. 2.3, with Theorem 1.4.22 and a new result, due to Lieb and Steiringer [112], on the concavity of a certain trace function on the cone of positive-definite matrices.

Theorem 1.4.22 allows us to relate the behavior of the $k$th eigenvalue of a matrix to the behavior of the largest eigenvalue of an appropriate compression of the matrix.

**Theorem 2.13.1 (Tropp [111]).** *Let* $\mathbf{X}$ *be a random, Hermitian matrix with dimension* $n$*, and let* $k \le n$ *be an integer. Then, for all* $t \in \mathbb{R}$*,*

$$\mathbb{P}\left(\lambda_k\left(\mathbf{X}\right) \geqslant t\right) \leqslant \inf_{\theta > 0} \min_{\mathbf{V} \in \mathbb{V}_{n-k+1}^n} \left\{ e^{-\theta t} \cdot \mathbb{E}\,\mathrm{Tr}\, e^{\theta \mathbf{V}^* \mathbf{X} \mathbf{V}} \right\}. \qquad (2.65)$$

*Proof.* Let $\theta$ be a fixed positive number. Then

$$\begin{aligned} \mathbb{P}\left(\lambda_k\left(\mathbf{X}\right) \geqslant t\right) = \mathbb{P}\left(\lambda_k\left(\theta\mathbf{X}\right) \geqslant \theta t\right) &= \mathbb{P}\left(e^{\lambda_k(\theta\mathbf{X})} \geqslant e^{\theta t}\right) \\ &\leqslant e^{-\theta t} \cdot \mathbb{E} e^{\lambda_k(\theta\mathbf{X})} \\ &= e^{-\theta t} \cdot \mathbb{E}\exp\left\{ \min_{\mathbf{V} \in \mathbb{V}_{n-k+1}^n} \lambda_{\max}\left(\theta \mathbf{V}^* \mathbf{X} \mathbf{V}\right) \right\}. \end{aligned}$$

The first identify follows from the positive homogeneity of eigenvalue maps and the second from the monotonicity of the scalar exponential function. The final two steps are Markov's inequality and (1.89).

Let us bound the expectation. Interchange the order of the exponential and the minimum, due to the monotonicity of the scalar exponential function; then apply the spectral mapping Theorem 1.4.4 to see that

$$\begin{aligned}\mathbb{E}\exp\left\{\min_{\mathbf{V}\in\mathbb{V}_{n-k+1}^{n}}\lambda_{\max}\left(\theta\mathbf{V}^{*}\mathbf{X}\mathbf{V}\right)\right\} &= \mathbb{E}\min_{\mathbf{V}\in\mathbb{V}_{n-k+1}^{n}}\lambda_{\max}\left(\exp\left(\theta\mathbf{V}^{*}\mathbf{X}\mathbf{V}\right)\right)\\ &\leqslant \min_{\mathbf{V}\in\mathbb{V}_{n-k+1}^{n}}\mathbb{E}\lambda_{\max}\left(\exp\left(\theta\mathbf{V}^{*}\mathbf{X}\mathbf{V}\right)\right)\\ &\leqslant \min_{\mathbf{V}\in\mathbb{V}_{n-k+1}^{n}}\mathbb{E}\,\mathrm{Tr}\left(\exp\left(\theta\mathbf{V}^{*}\mathbf{X}\mathbf{V}\right)\right).\end{aligned}$$

The first step uses Jensen's inequality. The second inequality follows because the exponential of a Hermitian matrix is always positive definite—see Sect. 1.4.16, so its largest eigenvalue is smaller than its trace. The trace functional is linear, which is very critical. The expectation is also linear. Thus we can exchange the order of the expectation and the trace: trace and expectation commute—see (1.87).

Combine these observations and take the infimum over all positive $\theta$ to complete the argument. □

Now let apply Theorem 2.13.1 to the case that $\mathbf{X}$ can be expressed as a sum of independent, Hermitian, random matrices. In this case, we develop the right-hand side of the Laplace transform bound (2.65) by using the following result.

**Theorem 2.13.2 (Tropp [111]).** *Consider a finite sequence $\{\mathbf{X}_i\}$ of independent, Hermitian, random matrices with dimension $n$ and a sequence $\{\mathbf{A}_i\}$ of **fixed** Hermitian matrices with dimension $n$ that satisfy the relations*

$$\mathbb{E}\left(e^{\mathbf{X}_i}\right)\leqslant e^{\mathbf{A}_i}. \tag{2.66}$$

*Let $\mathbf{V}\in\mathbb{V}_k^n$ be an isometric embedding of $\mathbb{C}^k$ into $\mathbb{C}^n$ for some $k\leq n$. Then*

$$\mathbb{E}\,\mathrm{Tr}\exp\left\{\sum_i\mathbf{V}^*\mathbf{X}_i\mathbf{V}\right\}\leqslant\mathrm{Tr}\exp\left\{\sum_i\mathbf{V}^*\mathbf{A}_i\mathbf{V}\right\}. \tag{2.67}$$

*In particular,*

$$\mathbb{E}\,\mathrm{Tr}\exp\left\{\sum_i\mathbf{X}_i\right\}\leqslant\mathrm{Tr}\exp\left\{\sum_i\mathbf{A}_i\right\}. \tag{2.68}$$

Theorem 2.13.2 is an extension of Lemma 2.5.1, which establish the result of (2.68). The proof depends on a recent result of [112], which extends Lieb's earlier classical result [50, Theorem 6]. Here $\mathbb{M}_{\mathrm{H}}^n$ represents the set of Hermitian matrices of $n\times n$.

**Proposition 2.13.3 (Lieb-Seiringer 2005).** *Let $\mathbf{H}$ be a Hermitian matrix with dimension $k$. Let $\mathbf{V}\in\mathbb{V}_k^n$ be an isometric embedding of $\mathbb{C}^k$ into $\mathbb{C}^n$ for some $k\leq n$. Then the function*

$$\mathbf{A}\mapsto\mathrm{Tr}\exp\left\{\mathbf{H}+\mathbf{V}^*\left(\log\mathbf{A}\right)\mathbf{V}\right\}$$

*is concave on the cone of positive-definite matrices in $\mathbb{M}_H^n$.*

*Proof of Theorem 2.13.2.* First, combining the given condition (2.66) with the operator monotonicity of the matrix logarithm gives the following for each $k$:

$$\log \mathbb{E} e^{\mathbf{X}_k} \leqslant \mathbf{A}_k. \tag{2.69}$$

Let $\mathbb{E}_k$ denote the expectation conditioned on the first $k$ summands, $\mathbf{X}_1$ through $\mathbf{X}_k$. Then

$$\begin{aligned}
\mathbb{E}\,\mathrm{Tr}\exp\left\{\sum_{i\leqslant j}\mathbf{V}^*\mathbf{X}_i\mathbf{V}\right\} &= \mathbb{E}\mathbb{E}_1\cdots\mathbb{E}_{j-1}\,\mathrm{Tr}\exp\left\{\sum_{i\leqslant j-1}\mathbf{V}^*\mathbf{X}_i\mathbf{V}+\mathbf{V}^*\left(\log e^{\mathbf{X}_j}\right)\mathbf{V}\right\}\\
&\leqslant \mathbb{E}\mathbb{E}_1\cdots\mathbb{E}_{j-2}\,\mathrm{Tr}\exp\left\{\sum_{i\leqslant j-1}\mathbf{V}^*\mathbf{X}_i\mathbf{V}+\mathbf{V}^*\left(\log \mathbb{E}e^{\mathbf{X}_j}\right)\mathbf{V}\right\}\\
&\leqslant \mathbb{E}\mathbb{E}_1\cdots\mathbb{E}_{j-2}\,\mathrm{Tr}\exp\left\{\sum_{i\leqslant j-1}\mathbf{V}^*\mathbf{X}_i\mathbf{V}+\mathbf{V}^*\left(\log e^{\mathbf{A}_j}\right)\mathbf{V}\right\}\\
&= \mathbb{E}\mathbb{E}_1\cdots\mathbb{E}_{j-2}\,\mathrm{Tr}\exp\left\{\sum_{i\leqslant j-1}\mathbf{V}^*\mathbf{X}_i\mathbf{V}+\mathbf{V}^*\mathbf{A}_j\mathbf{V}\right\}.
\end{aligned}$$

The first step follows from Proposition 2.13.3 and Jensen's inequality, and the second depends on (2.69) and the monotonicity of the trace exponential. Iterate this argument to complete the proof. The main result follows from combining Theorems 2.13.1 and 2.13.2. □

**Theorem 2.13.4 (Minimax Laplace Transform).** *Consider a finite sequence $\{\mathbf{X}_i\}$ of independent, random, Hermitian matrices with dimension $n$, and let $k \le n$ be an integer.*

1. *Let $\{\mathbf{A}_i\}$ be a sequence of Hermitian matrices that satisfy the semidefinite relations*

$$\mathbb{E}\left(e^{\theta\mathbf{X}_i}\right) \leqslant e^{g(\theta)\mathbf{A}_i}$$

*where $g:(0,\infty)\to[0,\infty)$. Then, for all $t\in\mathbb{R}$,*

$$\mathbb{P}\left(\lambda_k\left(\sum_i \mathbf{X}_i\right)\geqslant t\right) \leqslant \inf_{\theta>0}\min_{\mathbf{V}\in\mathbb{V}_{n-k+1}^n}\left[e^{-\theta t}\cdot\mathrm{Tr}\exp\left\{g(\theta)\sum_i\mathbf{V}^*\mathbf{A}_i\mathbf{V}\right\}\right].$$

2. *$\mathbf{A}_i : \mathbb{V}_{n-k+1}^n \to \mathbb{M}_H^n$ be a sequence of functions that satisfy the semidefinite relations*

$$\mathbb{E}\left(e^{\theta\mathbf{V}^*\mathbf{X}_i\mathbf{V}}\right) \leqslant e^{g(\theta)\mathbf{A}_i(\mathbf{V})}$$

*for all $\mathbf{V}\in\mathbb{V}_{n-k+1}^n$ where $g:(0,\infty)\to[0,\infty)$. Then, for all $t\in\mathbb{R}$,*

$$\mathbb{P}\left(\lambda_k\left(\sum_i \mathbf{X}_i\right) \geqslant t\right) \leqslant \inf_{\theta>0} \min_{\mathbf{V}\in\mathbb{V}^n_{n-k+1}} \left[e^{-\theta t}\cdot \operatorname{Tr}\exp\left\{g\left(\theta\right)\sum_i \mathbf{A}_i\left(\mathbf{V}\right)\right\}\right].$$

The first bound in Theorem 2.13.4 requires less detailed information on how compression affects the summands but correspondingly does not yield as sharp results as the second.

## 2.14 Chernoff Bounds for Interior Eigenvalues

Classical Chernoff bounds in Sect. 1.1.4 establish that the tails of a sum of independent, nonnegative, random variables decay subexponentially. Tropp [53] develops Chernoff bounds for the maximum and minimum eigenvalues of a *sum of independent, positive-semidefinite matrices*. In particular, sample covariance matrices are positive-semidefinite and the sums of independent, sample covariance matrices are ubiquitous. Following Gittens and Tropp [111], we extend this analysis to study the interior eigenvalues. The analogy with the scalar-valued random variables in Sect. 1.1.4 is aimed at, in this development. At this point, it is insightful if the audience reviews the materials in Sects. 1.1.4 and 1.3.

Intuitively, how concentrated the summands will determine the eigenvalues tail bounds; in other words, if we align the ranges of some operators, the maximum eigenvalue of a sum of these operators varies probably more than that of a sum of operators whose ranges are orthogonal. We are interested in a finite sequence of random summands $\{\mathbf{X}_i\}$. This sequence will concentrate in a given subspace. To measure how much this sequence concentrate, we define a function $\psi : \cup_{1\leqslant k\leqslant n}\mathbb{V}^n_k \to \mathbb{R}$ that has the property

$$\max_i \lambda_{\max}\left(\mathbf{V}^*\mathbf{X}_i\mathbf{V}\right) \leqslant \psi\left(\mathbf{V}\right) \text{ almost surely for each } \mathbf{V}\in\cup_{1\leqslant k\leqslant n}\mathbb{V}^n_k. \quad (2.70)$$

**Theorem 2.14.1 (Eigenvalue Chernoff Bounds [111]).** *Consider a finite sequence $\{\mathbf{X}_i\}$ of independent, random, positive-semidefinite matrices with dimension $n$. Given an integer $k \le n$, define*

$$\mu_k = \lambda_k\left(\sum_i \mathbb{E}\mathbf{X}_i\right),$$

*and let $\mathbf{V}_+ \in \mathbb{V}^n_{n-k+1}$ and $\mathbf{V}_- \in \mathbb{V}^n_k$ be isometric embeddings that satisfy*

$$\mu_k = \lambda_{\max}\left(\sum_i \mathbf{V}_+^*\mathbb{E}\mathbf{X}_i\mathbf{V}_+\right) = \lambda_{\min}\left(\sum_i \mathbf{V}_-^*\mathbb{E}\mathbf{X}_i\mathbf{V}_-\right).$$

*Then*

$$\mathbb{P}\left(\lambda_k\left(\sum_i \mathbf{X}_i\right) \geqslant (1+\delta)\,\mu_k\right) \leqslant (n-k+1)\cdot\left[\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right]^{\mu_k/\psi(\mathbf{V}_+)} \quad \text{for } \delta>0, \text{ and}$$

$$\mathbb{P}\left(\lambda_k\left(\sum_i \mathbf{X}_i\right) \leqslant (1-\delta)\,\mu_k\right) \leqslant k\cdot\left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\mu_k/\psi(\mathbf{V}_-)} \quad \text{for } \delta\in[0,1],$$

*where $\psi$ is a function that satisfies* (2.70)*.*

Practically, if it is difficult to estimate $\psi(\mathbf{V}_+)$ and $\psi(\mathbf{V}_-)$, we can use the weaker estimates

$$\psi(\mathbf{V}_+) \leqslant \max_{\mathbf{V}\in\mathbb{V}^n_{n-k+1}} \max_i \|\mathbf{V}^*\mathbf{X}_i\mathbf{V}\| = \max_i \|\mathbf{X}_i\|$$

$$\psi(\mathbf{V}_-) \leqslant \max_{\mathbf{V}\in\mathbb{V}^n_{k}} \max_i \|\mathbf{V}^*\mathbf{X}_i\mathbf{V}\| = \max_i \|\mathbf{X}_i\|.$$

The following lemma is due to Ahlswede and Winter [36]; see also [53, Lemma 5.8].

**Lemma 2.14.2.** *Suppose that $\mathbf{X}$ is a random positive-semidefinite matrix that satisfies $\lambda_{\max}(\mathbf{X}) \leqslant 1$. Then*

$$\mathbb{E}e^{\theta\mathbf{X}} \leqslant \exp\left(\left(e^{\theta}-1\right)(\mathbb{E}\mathbf{X})\right) \text{ for } \theta\in\mathbb{R}.$$

*Proof of Theorem 2.14.1, upper bound.* Without loss of generality, we consider the case $\psi(\mathbf{V}_+)=1$; the general case follows due to homogeneity. Define

$$\mathbf{A}_i(\mathbf{V}_+) = \mathbf{V}_+^*\mathbb{E}\mathbf{X}_i\mathbf{V}_+ \text{ and } g(\theta)=e^{\theta}-1.$$

Using Theorem 2.13.4 and Lemma 2.14.2 gives

$$\mathbb{P}\left(\lambda_k\left(\sum_i \mathbf{X}_i\right) \geqslant (1+\delta)\,\mu_k\right) \leqslant \inf_{\theta>0} e^{-\theta(1+\delta)\mu_k}\cdot \operatorname{Tr}\exp\left\{g(\theta)\sum_i \mathbf{V}_+^*\mathbb{E}\mathbf{X}_i\mathbf{V}_+\right\}.$$

The trace can be bounded by the maximum eigenvalue (since the maximum eigenvalue is nonnegative), by taking into account the reduced dimension of the summands:

$$\begin{aligned}\operatorname{Tr}\exp\left\{g(\theta)\sum_i \mathbf{V}_+^*\mathbb{E}\mathbf{X}_i\mathbf{V}_+\right\} &\leqslant (n-k+1)\cdot\lambda_{\max}\left(\exp\left\{g(\theta)\sum_i \mathbf{V}_+^*\mathbb{E}\mathbf{X}_i\mathbf{V}_+\right\}\right)\\ &= (n-k+1)\cdot\exp\left\{g(\theta)\cdot\lambda_{\max}\left(\sum_i \mathbf{V}_+^*\mathbb{E}\mathbf{X}_i\mathbf{V}_+\right)\right\}.\end{aligned}$$

The equality follows from the spectral mapping theorem (Theorem 1.4.4 at Page 34). We identify the quantity $\mu_k$; then combine the last two inequalities to give

$$\mathbb{P}\left(\lambda_k\left(\sum_i \mathbf{X}_i\right) \geqslant (1+\delta)\,\mu_k\right) \leqslant (n-k+1)\cdot \inf_{\theta>0} e^{[g(\theta)-\theta(1+\delta)]\mu_k}.$$

By choosing $\theta = \log(1+\delta)$, the right-hand side is minimized (by taking care of the infimum), which gives the desired upper tail bound. $\square$

*Proof of Theorem 2.14.1, lower bound.* The proof of lower bound is very similar to that of upper bound above. As above, consider only $\psi(\mathbf{V}_-) = 1$. It follow from (1.91) (Page 47) that

$$\mathbb{P}\left(\lambda_k\left(\sum_i \mathbf{X}_i\right) \leqslant (1-\delta)\,\mu_k\right) = \mathbb{P}\left(\lambda_{n-k+1}\left(\sum_i -\mathbf{X}_i\right) \geqslant -(1-\delta)\,\mu_k\right). \tag{2.71}$$

Applying Lemma 2.14.2, we find that, for $\theta > 0$,

$$\begin{aligned}\mathbb{E}e^{\theta\left(-\mathbf{V}_{-^*}\mathbf{x}_i\mathbf{v}_-\right)} = \mathbb{E}e^{(-\theta)\mathbf{V}_-^*\mathbf{X}_i\mathbf{V}_-} &\leqslant \exp\left(g(\theta)\cdot\left(\mathbb{E}\left[-\mathbf{V}_-^*\mathbf{X}_i\mathbf{V}_-\right]\right)\right)\\ &= \exp\left(g(\theta)\cdot\left(\mathbf{V}_-^*\left(-\mathbb{E}\mathbf{X}_i\right)\mathbf{V}_-\right)\right),\end{aligned}$$

where $g(\theta) = 1-e^{\theta}$. The last equality follows from the linearity of the expectation. Using Theorem 2.13.4, we find the latter probability in (2.71) is bounded by

$$\inf_{\theta>0} e^{\theta(1-\delta)\mu_k}\cdot \operatorname{Tr}\exp\left\{g(\theta)\sum_i \mathbf{V}_-^*\left(-\mathbb{E}\mathbf{X}_i\right)\mathbf{V}_-\right\}.$$

The trace can be bounded by the maximum eigenvalue (since the maximum eigenvalue is nonnegative), by taking into account the reduced dimension of the summands:

$$\begin{aligned}\operatorname{Tr}\exp\left\{g(\theta)\sum_i \mathbf{V}_-^*\left(-\mathbb{E}\mathbf{X}_i\right)\mathbf{V}_-\right\} &\leqslant k\cdot\lambda_{\max}\left(\exp\left\{g(\theta)\sum_i \mathbf{V}_-^*\left(-\mathbb{E}\mathbf{X}_i\right)\mathbf{V}_-\right\}\right)\\ &= k\cdot\exp\left\{-g(\theta)\cdot\lambda_{\min}\left(\sum_i \mathbf{V}_-^*\left(\mathbb{E}\mathbf{X}_i\right)\mathbf{V}_-\right)\right\}\\ &= k\cdot\exp\left\{-g(\theta)\cdot\mu_k\right\}.\end{aligned}$$

The equality follows from the spectral mapping theorem, Theorem 1.4.4 (Page 34), and (1.92) (Page 47). In the second equality, we identify the quantity $\mu_k$. Note that $-g(\theta) \leq 0$. Our argument establishes the bound

$$\mathbb{P}\left(\lambda_k\left(\sum_i \mathbf{X}_i\right) \geqslant (1+\delta)\,\mu_k\right) \leqslant k\cdot \inf_{\theta>0} e^{[\theta(1+\delta)-g(\theta)]\mu_k}.$$

The right-hand side is minimized, (by taking care of the infimum), when $\theta = -\log(1-\delta)$, which gives the desired upper tail bound. $\square$

From the two proofs, we see the property that the maximum eigenvalue is nonnegative is fundamental. Using this property, we convert the trace functional into the maximum eigenvalue functional. Then the Courant-Fischer theorem, Theorem 1.4.22, can be used. The spectral mapping theorem is applied almost everywhere; it is must be recalled behind the mind. The non-commutative property is fundamental in studying random matrices. By using the eigenvalues and their variation property, it is very convenient to think of random matrices as scalar-valued random variables, in which we convert the two dimensional problem into one-dimensional problem—much more convenient to handle.

## 2.15 Linear Filtering Through Sums of Random Matrices

The linearity of the expectation and the trace is so basic. We must always bear this mind. The trace which is a linear functional converts a random matrix into a scalar-valued random variable; so as the $k$th interior eigenvalue which is a non-linear functional. Since trace and expectation commute, it follows from (1.87), which says that

$$\mathbb{E}\left(\operatorname{Tr}\mathbf{X}\right) = \operatorname{Tr}\left(\mathbb{E}\mathbf{X}\right). \tag{2.72}$$

As said above, in the left-hand side, $\operatorname{Tr}\mathbf{X}$ is a scalar-valued random variable, so its expectation is treated as our standard textbooks on random variables and processes; remarkably, in the right-hand side, the expectation of a random matrix $\mathbb{E}\mathbf{X}$ is also a matrix whose entries are expected values. After this expectation, a trace functional converts the matrix value into a scalar value. One cannot help replacing $\mathbb{E}\mathbf{X}$ with the empirical average—a sum of random matrices, that is,

$$\mathbb{E}\mathbf{X} \cong \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i, \tag{2.73}$$

as we deal with the scalar-valued random variables. This intuition lies at the very basis of modern probability. In this book, one purpose is to prepare us for the intuition of this "approximation" (2.73), for a given $n$, *large but finite*—the $n$ is taken as it is. We are not interested in the asymptotic limit as $n \to \infty$, rather the non-asymptotic analysis. One natural metric of measure is the $k$th interior eigenvalues

$$\lambda_k\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i - \mathbb{E}\mathbf{X}\right).$$

Note the interior eigenvalues are non-linear functionals. We cannot simply separate the two terms.

We can use the linear trace functional that is the sum of all eigenvalues. As a result, we have

$$\sum_k \lambda_k \left( \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i - \mathbb{E}\mathbf{X} \right) = \mathrm{Tr} \left( \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i - \mathbb{E}\mathbf{X} \right) = \mathrm{Tr} \left( \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \right) - \mathrm{Tr} \left( \mathbb{E}\mathbf{X} \right)$$
$$= \frac{1}{n} \sum_{i=1}^{n} \mathrm{Tr}\, \mathbf{X}_i - \mathbb{E} \left( \mathrm{Tr}\, \mathbf{X} \right).$$

The linearity of the trace is used in the second and third equality. The property that trace and expectation commute is used in the third equality. Indeed, the linear trace functional is convenient, but a lot of statistical information is contained in the interior eigenvalues. For example, the median of the eigenvalues, rather than the average of the eigenvalues—the trace divided by its dimension can be viewed as the average, is more representative statistically.

We are in particular interested in the signal plus noise model in the matrix setting. We consider instead

$$\mathbb{E} \left( \mathbf{X} + \mathbf{Z} \right) \cong \frac{1}{n} \sum_{i=1}^{n} \left( \mathbf{X}_i + \mathbf{Z}_i \right), \tag{2.74}$$

for

$$\mathbf{X}, \mathbf{Z}, \mathbf{X}_i, \mathbf{Z}_i \geqslant 0 \quad \text{and} \quad \mathbf{X}, \mathbf{Z}, \mathbf{X}_i, \mathbf{Z}_i \in \mathbb{C}^{m \times m},$$

where $\mathbf{X}, \mathbf{X}_i$ represent the signal and $\mathbf{Z}, \mathbf{Z}_i$ the noise. Recall that $\mathbf{A} \geq \mathbf{0}$ means that $\mathbf{A}$ is positive semidefinite (Hermitian and all eigenvalues of $\mathbf{A}$ are nonnegative). Samples covariance matrices of dimensions $m \times m$ are most often used in this context.

Since we have a prior knowledge that $\mathbf{X}, \mathbf{X}_i$ are of low rank, the low-rank matrix recovery naturally fits into this framework. We can choose the matrix dimension $m$ such that enough information of the signal matrices $\mathbf{X}_i$ is recovered, but we don't care if sufficient information of $\mathbf{Z}_i$ can be recovered for this chosen $m$. For example, only the first dominant $k$ eigenvalues of $\mathbf{X}, \mathbf{X}_i$ are recovered, which will be treated in Sect. 2.10 Low Rank Approximation. We conclude that the sums of random matrices have the fundamental nature of imposing the structures of the data that only exhibit themselves in the matrix setting. The low rank and the positive semi-definite of sample covariance matrices belong to these data structures. When the data is big, we must impose these additional structures for high-dimensional data processing.

The intuition of exploiting (2.74) is as follows: if the estimates of $\mathbf{X}_i$ are so accurate that they are independent and identically distributed $\mathbf{X}_i = \mathbf{X}_0$, then we rewrite (2.74) as

$$\mathbb{E} \left( \mathbf{X} + \mathbf{Z} \right) \cong \frac{1}{n} \sum_{i=1}^{n} \left( \mathbf{X}_i + \mathbf{Z}_i \right) = \frac{1}{n} \sum_{i=1}^{n} \left( \mathbf{X}_0 + \mathbf{Z}_i \right) = \mathbf{X}_0 + \frac{1}{n} \sum_{i=1}^{n} \mathbf{Z}_i. \tag{2.75}$$

Practically, $\mathbf{X} + \mathbf{Z}$ cannot be separated. The exploitation of the additional low-rank structure of the signal matrices allows us to extract the signal matrix $\mathbf{X}_0$. The average of the noise matrices $\frac{1}{n}\sum_{i=1}^{n} \mathbf{Z}_i$ will reduce the total noise power (total variance), while the signal power is kept constant. This process can effectively improve the signal to noise ratio, which is especially critical to detection of extremely weak signals (relative to the noise power).

The basic observation is that the above data processing only involves the **linear operations**. The process is also blind, which means that no prior knowledge of the noise matrix is used. We only take advantage of the rank structure of the underlying signal and noise matrices: the dimensions of the signal space is lower than the dimensions of the noise space. The above process can be extended to more general case: $\mathbf{X}_i$ are more dependent than $\mathbf{Z}_i$, where $\mathbf{X}_i$ are dependent on each other and so are $\mathbf{Z}_i$ , but $\mathbf{X}_i$ are independent of $\mathbf{Z}_i$. Thus, we rewrite (2.74) as

$$\mathbb{E}\left(\mathbf{X} + \mathbf{Z}\right) \cong \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{X}_i + \mathbf{Z}_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i + \frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i. \tag{2.76}$$

All we care is that, through the sums of random matrices, $\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i$ is performing statistically better than $\frac{1}{n}\sum_{i=1}^{n} \mathbf{Z}_i$ . For example, we can use the linear trace functional (average operation) and the non-linear median functional. To calculate the median value of $\lambda_k, 1 \le k \le n$,

$$\mathbb{M}\left[\lambda_k\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i + \frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i\right)\right],$$

where $\mathbb{M}$ is the median value which is a scalar-valued random variable, we need to calculate

$$\lambda_k\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i + \frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i\right) \quad 1 \le k \le n.$$

The average operation comes down to a trace operation

$$\frac{1}{n}\sum_{i=1}^{n}\operatorname{Tr}\mathbf{X}_i + \frac{1}{n}\sum_{i=1}^{n}\operatorname{Tr}\mathbf{Z}_i,$$

where the linearity of the trace is used. This is simply the standard sum of scalar-valued random variables. It is expected, via the central limit theorem, that their sum approaches to the Gaussian distribution, for a reasonably large $n$. As pointed out before, this trace operation throws away a lot of statistical information that is available in the random matrices, for example, the matrix structures.

## 2.16 Dimension-Free Inequalities for Sums of Random Matrices

Sums of random matrices arise in many statistical and probabilistic applications, and hence their concentration behavior is of fundamental significance. Surprisingly, the classical exponential moment method used to derive tail inequalities for scalar random variables carries over to the matrix setting when augmented with certain matrix trace inequalities [68, 113]. Altogether, these results have proven invaluable in constructing and simplifying many probabilistic arguments concerning sums of random matrices.

One deficiency of many of these previous inequalities is their dependence on the explicit matrix dimension, which prevents their application to infinite dimensional spaces that arise in a variety of data analysis tasks, such as kernel based machine learning [114]. In this subsection, we follow [68, 113] to prove analogous results where dimension is replaced with a trace quantity that can be small, even when the explicit matrix dimension is large or infinite. Magen and Zouzias [115] also gives similar results that are complicated and fall short of giving an exponential tail inequality.

We use $\mathbb{E}_i\left[\cdot\right]$ as shorthand for $\mathbb{E}_i\left[\cdot\right] = \mathbb{E}_i\left[\cdot\left|\mathbf{X}_1, \ldots, \mathbf{X}_i\right.\right]$, the conditional expectation. The main idea is to use Theorem 1.4.17: Lieb's theorem.

**Lemma 2.16.1 (Tropp [49]).** *Let* $\mathbf{I}$ *be the identity matrix for the range of the* $\mathbf{X}_i$*. Then*

$$\mathbb{E}\left[\operatorname{Tr}\left(\exp\left(\sum_{i=1}^{N}\mathbf{X}_i - \sum_{i=1}^{N}\ln\mathbb{E}_i\left[\exp\left(\mathbf{X}_i\right)\right]\right) - \mathbf{I}\right)\right] \leqslant 0. \tag{2.77}$$

*Proof.* We follow [113]. The induction method is used for the proof. For $N = 0$, it is easy to check the lemma is correct. For $N \geq 1$, assume as the inductive hypothesis that (2.77) holds with $N$ with replaced with $N-1$. In this case, we have that

$$\begin{aligned}
&\mathbb{E}\left[\operatorname{Tr}\left(\exp\left(\sum_{i=1}^{N}\mathbf{X}_i - \sum_{i=1}^{N}\ln\mathbb{E}_i\left[\exp\left(\mathbf{X}_i\right)\right]\right) - \mathbf{I}\right)\right]\\
&= \mathbb{E}\left[\mathbb{E}_N\left[\operatorname{Tr}\left(\exp\left(\sum_{i=1}^{N-1}\mathbf{X}_i - \sum_{i=1}^{N}\ln\mathbb{E}_i\left[\exp\left(\mathbf{X}_i\right)\right] + \ln\exp\left(\mathbf{X}_N\right)\right) - \mathbf{I}\right)\right]\right]\\
&\leqslant \mathbb{E}\left[\mathbb{E}_N\left[\operatorname{Tr}\left(\exp\left(\sum_{i=1}^{N-1}\mathbf{X}_i - \sum_{i=1}^{N}\ln\mathbb{E}_i\left[\exp\left(\mathbf{X}_i\right)\right] + \ln\mathbb{E}_N\exp\left(\mathbf{X}_N\right)\right) - \mathbf{I}\right)\right]\right]\\
&= \mathbb{E}\left[\mathbb{E}_N\left[\operatorname{Tr}\left(\exp\left(\sum_{i=1}^{N-1}\mathbf{X}_i - \sum_{i=1}^{N-1}\ln\mathbb{E}_i\left[\exp\left(\mathbf{X}_i\right)\right]\right) - \mathbf{I}\right)\right]\right]\\
&\leqslant 0
\end{aligned}$$

where the second line follows from Theorem 1.4.17 and Jensen's inequality. The fifth line follows from the inductive hypothesis. □

While (2.77) gives the trace result, sometimes we need the largest eigenvalue.

**Theorem 2.16.2 (Largest eigenvalue—Hsu, Kakade and Zhang [113]).** *For any* $\alpha \in \mathbb{R}$ *and any* $t > 0$

$$\mathbb{P}\left[\lambda_{\max}\left(\alpha\sum_{i=1}^{N}\mathbf{X}_i - \sum_{i=1}^{N}\log\mathbb{E}_i\left[\exp\left(\alpha\mathbf{X}_i\right)\right]\right) > t\right] \leqslant \operatorname{Tr}\left(\mathbb{E}\left(-\alpha\sum_{i=1}^{N}\mathbf{X}_i + \sum_{i=1}^{N}\log\mathbb{E}_i\left[\exp\left(\alpha\mathbf{X}_i\right)\right]\right)\right)\cdot\left(e^{t} - t - 1\right)^{-1}.$$

*Proof.* Define a new matrix $\mathbf{A} = \alpha\sum_{i=1}^{N}\mathbf{X}_i - \sum_{i=1}^{N}\log\mathbb{E}_i\left[\exp\left(\alpha\mathbf{X}_i\right)\right]$.. Note that $g(x) = e^x - x - 1$ is non-negative for all real $x$, and increasing for $x \geq 0$. Let $\lambda_i\left(\mathbf{A}\right)$ be the $i$-th eigenvalue of the matrix $\mathbf{A}$, we have

$$\begin{aligned}
\mathbb{P}\left[\lambda_{\max}\left(\mathbf{A}\right) > t\right]\left(e^t - t - 1\right) &= \mathbb{E}\left[\mathcal{I}\left(\lambda_{\max}\left(\mathbf{A}\right) > t\right)\left(e^t - t - 1\right)\right] \\
&\leqslant \mathbb{E}\left(e^{\lambda_{\max}(\mathbf{A})} - \lambda_{\max}\left(\mathbf{A}\right) - 1\right) \\
&\leqslant \mathbb{E}\left(\sum_{i}\left(e^{\lambda_i(\mathbf{A})} - \lambda_i\left(\mathbf{A}\right) - 1\right)\right) \\
&\leqslant \mathbb{E}\left(\operatorname{Tr}\left[\exp\left(\mathbf{A}\right) - \mathbf{A} - \mathbf{I}\right]\right) \\
&\leqslant \operatorname{Tr}\left(\mathbb{E}\left[-\mathbf{A}\right]\right)
\end{aligned}$$

where $\mathcal{I}(x)$ is the indicator function of $x$. The second line follows from the spectral mapping theorem. The third line follows from the increasing property of the function $g(x)$. The last line follows from Lemma 2.16.1. □

When $\sum_{i=1}^{N}\mathbf{X}_i$ is zero mean, then the first term in Theorem 2.16.2 vanishes, so the trace term

$$\operatorname{Tr}\left(\mathbb{E}\left(\sum_{i=1}^{N}\log\mathbb{E}_i\left[\exp\left(\alpha\mathbf{X}_i\right)\right]\right)\right)$$

can be made small by an appropriate choice of $\alpha$.

**Theorem 2.16.3 (Matrix sub-Gaussian bound—Hsu, Kakade and Zhang [113]).** *If there exists* $\bar{\sigma} > 0$ *and* $\bar{\kappa} > 0$ *such that for all* $i = 1, \ldots, N$,

$$
\begin{aligned}
&\mathbb{E}_i\left[\mathbf{X}_i\right]=0 \\
&\lambda_{\max}\left(\frac{1}{N}\sum_{i=1}^{N}\log\mathbb{E}_i\left[\exp\left(\alpha\mathbf{X}_i\right)\right]\right)\leqslant\frac{\alpha^2\bar{\sigma}^2}{2} \\
&\mathbb{E}\left(\operatorname{Tr}\left(\frac{1}{N}\sum_{i=1}^{N}\log\mathbb{E}_i\left[\exp\left(\alpha\mathbf{X}_i\right)\right]\right)\right)\leqslant\frac{\alpha^2\bar{\sigma}^2\bar{\kappa}}{2}
\end{aligned}
$$

*for all* $\alpha > 0$ *almost surely, then for any* $t > 0$*,*

$$
\mathbb{P}\left[\lambda_{\max}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i\right)>\sqrt{\frac{2\bar{\sigma}^2 t}{N}}\right]\leqslant\bar{\kappa}\cdot t\big(e^t-t-1\big)^{-1}.
$$

The proof is simple. We refer to [113] for a proof.

**Theorem 2.16.4 (Matrix Bernstein Bound—Hsu, Kakade and Zhang [113]).** *If there exists* $\bar{b} > 0,\ \bar{\sigma} > 0$ *and* $\bar{\kappa} > 0$ *such that for all* $i = 1, \ldots, N$*,*

$$
\begin{aligned}
&\mathbb{E}_i\left[\mathbf{X}_i\right]=0 \\
&\lambda_{\max}\left(\mathbf{X}_i\right)\leqslant\bar{b} \\
&\lambda_{\max}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_i\left[\mathbf{X}_i^2\right]\right)\leqslant\bar{\sigma}^2 \\
&\mathbb{E}\left(\operatorname{Tr}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_i\left[\mathbf{X}_i^2\right]\right)\right)\leqslant\bar{\sigma}^2\bar{\kappa}
\end{aligned}
$$

*almost surely, then for any* $t > 0$*,*

$$
\mathbb{P}\left[\lambda_{\max}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i\right)>\sqrt{\frac{2\bar{\sigma}^2 t}{N}}+\frac{\bar{b}t}{3N}\right]\leqslant\bar{\kappa}\cdot t\big(e^t-t-1\big)^{-1}.
$$

The proof is simple. We refer to [113] for a proof.

Explicit dependence on the dimension of the matrix does not allow straightforward use of these results in the infinite-dimensional setting. Minsker [116] deals with this issue by extension of previous results. This new result is of interest to low rank matrix recovery and approximation matrix multiplication. Let $||\cdot||$ denote the operator norm $\|\mathbf{A}\| = \max_i \{\lambda_i\left(\mathbf{A}\right)\}$, where $\lambda_i$ are eigenvalues of a Hermitian operator $\mathbf{A}$. Expectation $\mathbb{E}\mathbf{X}$ is taken elementwise.

**Theorem 2.16.5 (Dimension-free Bernstein inequality—Minsker [116]).** *Let* $\mathbf{X}_1, \ldots, \mathbf{X}_N$ *be a sequence of* $n \times n$ *independent Hermitian random matrices such* *that* $\mathbb{E}\mathbf{X}_i = 0$ *and* $||\mathbf{X}_i|| \le 1$ *almost surely. Denote* $\sigma^2 = \left\| \sum_{i=1}^{N} \mathbb{E}\mathbf{X}_i^2 \right\|$*. Then, for any* $t > 0$

$$\mathbb{P}\left( \left\| \sum_{i=1}^{N} \mathbf{X}_i \right\| > t \right) \leqslant 2 \frac{\mathrm{Tr}\left( \sum_{i=1}^{N} \mathbb{E}\mathbf{X}_i^2 \right)}{\sigma^2} \exp\left(-\Psi_\sigma(t)\right) \cdot r_\sigma(t)$$

*where* $\Psi_\sigma(t) = \frac{t^2/2}{\sigma^2 + t/3}$ *and* $r_\sigma(t) = 1 + \frac{6}{t^2 \log^2(1+t/\sigma^2)}$.

If $\sum_{i=1}^{N} \mathbb{E}\mathbf{X}_i^2$ is of approximately low rank, i.e., has many small eigenvalues, the number of non-zero eigenvalues are big. The term $\frac{\mathrm{Tr}\left( \sum_{i=1}^{N} \mathbb{E}\mathbf{X}_i^2 \right)}{\sigma^2}$, however, is can be much smaller than the dimension $n$. Minsker [116] has applied Theorem 2.16.5 to the problem of learning the continuous-time kernel.

A concentration inequality for the sums of matrix-valued martingale differences is also obtained by Minsker [116]. Let $\mathbb{E}_{i-1}[\cdot]$ stand for the conditional expectation $\mathbb{E}_{i-1}[\cdot\,|\mathbf{X}_1, \ldots, \mathbf{X}_i]$.

**Theorem 2.16.6 (Minsker [116]).** *Let* $\mathbf{X}_1, \ldots, \mathbf{X}_N$ *be a sequence of martingale differences with values in the set of* $n \times n$ *independent Hermitian random matrices such that* $||\mathbf{X}_i|| \le 1$ *almost surely. Denote* $\mathbf{W}_N = \sum_{i=1}^{N} \mathbb{E}_{i-1}\mathbf{X}_i^2$*. Then, for any* $t > 0$*,*

$$\mathbb{P}\left( \left\| \sum_{i=1}^{N} \mathbf{X}_i \right\| > t, \lambda_{\max}(\mathbf{W}_N) \leqslant \sigma^2 \right) \leqslant 2\mathrm{Tr}\left[ p\left( -\frac{t}{\sigma^2} \mathbb{E}\mathbf{W}_N \right) \right] \exp\left(-\Psi_\sigma(t)\right) \cdot \left( 1 + \frac{6}{\Psi_\sigma^2(t)} \right),$$

*where* $p(t) = \min(-t, 1)$.

## 2.17 Some Khintchine-Type Inequalities

**Theorem 2.17.1 (Non-commutative Bernstein-type inequality [53]).** *Consider a finite sequence* $\mathbf{X}_i$ *of independent centered Hermitian random* $n \times n$ *matrices. Assume we have for some numbers* $K$ *and* $\sigma$ *such that*

$$\|\mathbf{X}_i\| \leqslant K \quad \textit{almost surely,} \qquad \left\| \sum_i \mathbb{E}\mathbf{X}_i^2 \right\| \leqslant \sigma^2.$$

*Then, for every* $t \ge 0$*, we have*

$$\mathbb{P}\left(\|\mathbf{X}_i\| \geqslant t\right) \leqslant 2n \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + Kt/3}\right).$$

We say $\xi_1, \ldots, \xi_N$ are independent Bernoulli random variables when each $\xi_i$ takes on the values $\pm 1$ with equal probability. In 1923, in an effort to provide a sharp estimate on the rate of convergence in Borel's strong law of large numbers, Khintchine proved the following inequality that now bears his name.

**Theorem 2.17.2 (Khintchine's inequality [117]).** *Let $\xi_1, \ldots, \xi_N$ be a sequence of independent Bernoulli random variables, and let $X_1, \ldots, X_N$ be an arbitrary sequence of scalars. Then, for any $N = 1, 2, \ldots$ and $p \in (0, \infty)$, there exists an absolute constant $C_p > 0$ such that*

$$\mathbb{E}\left[\left|\sum_{i=1}^{N} \xi_i X_i\right|^p\right] \leqslant C_p \cdot \left(\sum_{i=1}^{N} |X_i|^2\right)^{p/2}. \tag{2.78}$$

In fact, Khintchine only established the inequality for the case where $p \geq 2$ is an even integer. Since his work, much effort has been spent on determining the optimal value of $C_p$ in (2.78). In particular, it has been shown [117] that for $p \geq 2$, the value

$$C_p^* = \left(\frac{2^p}{\pi}\right)^{1/2} \Gamma\left(\frac{p+1}{2}\right)$$

is the best possible. Here $\Gamma(\cdot)$ is the Gamma function. Using Stirling's formula, one can show [117] that $C_p^*$ is of the order $p^{p/2}$ for all $p \geq 2$.

The Khintchine inequality is extended to the case for arbitrary $m \times n$ matrices. Here $\|\mathbf{A}\|_{S_p}$ denotes the Schatten $p$-norm of an $m \times n$ matrix $\mathbf{A}$, i.e., $\|\mathbf{A}\|_{S_p} = \|\boldsymbol{\sigma}(\mathbf{A})\|_p$, where $\boldsymbol{\sigma} \in \mathbb{R}^{\min\{m,n\}}$ is the vector of singular values of $\mathbf{A}$, and $||\cdot||_p$ is the usual $l_p$-norm.

**Theorem 2.17.3 (Khintchine's inequality for arbitrary $m \times n$ matrices [118]).** *Let $\xi_1, \ldots, \xi_N$ be a sequence of independent Bernoulli random variables, and let $\mathbf{X}_1, \ldots, \mathbf{X}_N$ be arbitrary $m \times n$ matrices. Then, for any $N = 1, 2, \ldots$ and $p \geq 2$, we have*

$$\mathbb{E}\left[\left\|\sum_{i=1}^{N} \xi_i \mathbf{X}_i\right\|_{S_p}^p\right] \leqslant p^{p/2} \cdot \left(\sum_{i=1}^{N} \|\mathbf{X}_i\|_{S_p}^2\right)^{p/2}.$$

The normalization $\sum_{i=1}^{N} \|\mathbf{X}_i\|_{S_p}^2$ is not the only one possible in order for a Khintchine-type inequality to hold. In 1986, Lust-Piquard showed another one possibility.

**Theorem 2.17.4 (Non-Commutative Khintchine's inequality for arbitrary $m \times n$ matrices [119]).** *Let $\xi_1, \ldots, \xi_N$ be a sequence of independent Bernoulli random variables, and let $\mathbf{X}_1, \ldots, \mathbf{X}_N$ be an arbitrary sequence of $m \times n$ matrices. Then, for any $N = 1, 2, \ldots$ and $p \geq 2$, there exists an absolute constant $\gamma_p > 0$ such that*

$$\mathbb{E}\left[\left\|\sum_{i=1}^{N}\xi_i\mathbf{X}_i\right\|_{S_p}^{p}\right] \leqslant \gamma_p \cdot \max\left\{\left\|\left(\sum_{i=1}^{N}\mathbf{X}_i\mathbf{X}_i^T\right)^{1/2}\right\|_{S_p}^{p}, \left\|\left(\sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{X}_i\right)^{1/2}\right\|_{S_p}^{p}\right\}.$$

The proof of Lust-Piquard does not provide an estimate for $\gamma_p$. In 1998, Pisier [120] showed that

$$\gamma_p \leqslant \alpha p^{p/2}$$

for some absolute constant $\alpha > 0$. Using the result of Buchholz [121], we have

$$\alpha \leqslant (\pi/e)^{p/2}/2^{p/4} < 1$$

for all $p \geq 2$. We note that Theorem 2.17.4 is valid (with $\gamma_p \leqslant \alpha p^{p/2} < p^{p/2}$) when $\xi_1, \ldots, \xi_N$ are i.i.d. standard Gaussian random variables [121].

Let $\mathbf{C}_i$ be arbitrary $m \times n$ matrices such that

$$\sum_{i=1}^{N}\mathbf{C}_i\mathbf{C}_i^T \leqslant \mathbf{I}_m, \qquad \sum_{i=1}^{N}\mathbf{C}_i^T\mathbf{C}_i \leqslant \mathbf{I}_n. \tag{2.79}$$

So [122] derived another useful theorem.

**Theorem 2.17.5 (So [122]).** *Let $\xi_1, \ldots, \xi_N$ be independent mean zero random variables, each of which is either (i) supported on [$-1$,1], or (ii) Gaussian with variance one. Further, let $\mathbf{X}_1, \ldots, \mathbf{X}_N$ be arbitrary $m \times n$ matrices satisfying $\max(m, n) \geqslant 2$ and (2.79). Then, for any $t \geq 1/2$, we have*

$$\operatorname{Prob}\left(\left\|\sum_{i=1}^{N}\xi_i\mathbf{X}_i\right\| \geqslant \sqrt{2e\left(1+t\right)\ln\max\left\{m,n\right\}}\right) \leqslant \left(\max\left\{m,n\right\}\right)^{-t}$$

*if $\xi_1, \ldots, \xi_N$ are i.i.d. Bernoulli or standard normal random variables; and*

$$\operatorname{Prob}\left(\left\|\sum_{i=1}^{N}\xi_i\mathbf{X}_i\right\| \geqslant \sqrt{8e\left(1+t\right)\ln\max\left\{m,n\right\}}\right) \leqslant \left(\max\left\{m,n\right\}\right)^{-t}$$

*if $\xi_1, \ldots, \xi_N$ are independent mean zero random variables supported on [$-1$,1].*

We refer to Sect. 11.2 for a proof.

Here we state a result of [123] that is stronger than Theorem 2.17.4. We deal with bilinear form. Let $\mathbf{X}$ and $\mathbf{Y}$ be two matrices of size $n \times k_1$ and $n \times k_2$ which satisfy

$$\mathbf{X}^T\mathbf{Y} = 0$$

and let $\{\mathbf{x}_i\}$ and $\{\mathbf{y}_i\}$ be row vectors of $\mathbf{X}$ and $\mathbf{Y}$, respectively. Denote $\varepsilon_i$ to be a sequence of i.i.d. $\{0/1\}$ Bernoulli random variables with $\mathbb{P}(\varepsilon_i = 1) = \bar{\varepsilon}$. Then, for $p \geq 2$

$$\begin{aligned}\left(\mathbb{E}\left\|\sum_i \varepsilon_i \mathbf{x}_i^T \mathbf{y}_i\right\|_{S_p}^p\right)^{1/p} &\leqslant 2\sqrt{2}\gamma_p^2 \max_i \|\mathbf{x}_i\| \max_i \|\mathbf{y}_i\| + \\ 2\sqrt{\bar{\varepsilon}}\gamma_p \max &\left\{\max_i \|\mathbf{x}_i\| \left\|\sum_i \mathbf{y}_i^T \mathbf{y}_i\right\|_{S_p}^{1/2}, \max_i \|\mathbf{y}_i\| \left\|\sum_i \mathbf{x}_i^T \mathbf{x}_i\right\|_{S_p}^{1/2}\right\},\end{aligned} \tag{2.80}$$

where $\gamma_p$ is the absolute constant defined in Theorem 2.17.4. This proof of (2.80) uses the following result

$$\left(\mathbb{E}\left\|\sum_i \varepsilon_i \mathbf{x}_i^T \mathbf{x}_i\right\|_{S_p}^p\right)^{1/p} \leqslant 2\gamma_p^2 \max_i \|\mathbf{x}_i\|^2 + \bar{\varepsilon}\left\|\sum_i \mathbf{x}_i^T \mathbf{x}_i\right\|_{S_p}$$

for $p \geq 2$.

Now consider $\mathbf{X}^T\mathbf{X} = \mathbf{I}$, then for $p \geq \log k$, we have [123]

$$\left(\mathbb{E}\left\|\mathbf{I}_{k\times k} - \frac{1}{\bar{\varepsilon}}\sum_{i=1}^n \varepsilon_i \mathbf{x}_i^T \mathbf{x}_i\right\|_{S_p}^p\right)^{1/p} \leqslant C\sqrt{\frac{p}{\bar{\varepsilon}}} \max_i \|\mathbf{x}_i\|,$$

where $C = 2^{3/4}\sqrt{\pi e} \approx 5$. This result guarantees that the invertibility of a submatrix which is formed from sampling a few columns (or rows) of a matrix $\mathbf{X}$.

**Theorem 2.17.6 ([124]).** *Let $\mathbf{X} \in \mathbb{R}^{n\times n}$, be a random matrix whose entries are independent, zero-mean, random variables. Then, for $p \geq \log n$,*

$$\left(\mathbb{E}\|\mathbf{X}\|^p\right)^{1/p} \leqslant c_0 2^{1/p}\sqrt{p}\left(\sqrt{\mathbb{E}\left(\max_i \sum_j X_{ij}^2\right)^p} + \sqrt{\mathbb{E}\left(\max_j \sum_i X_{ij}^2\right)^p}\right)^{1/p},$$

*where $c_0 \leqslant 2^{3/4}\sqrt{\pi e} < 5$.*

**Theorem 2.17.7 ([124]).** *Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be any matrix and $\tilde{\mathbf{A}} \in \mathbb{R}^{n\times n}$ be a random matrix such that*

$$\mathbb{E}\tilde{\mathbf{A}} = \mathbf{A}.$$

*Then, for* $p \geq \log n$,

$$\left(\mathbb{E}\left\|\mathbf{A}-\tilde{\mathbf{A}}\right\|^p\right)^{1/p} \leqslant c_0 2^{1/p}\sqrt{p}\left(\sqrt{\mathbb{E}\left(\max_i \sum_j \tilde{A}_{ij}^2\right)^p} + \sqrt{\mathbb{E}\left(\max_j \sum_i \tilde{A}_{ij}^2\right)^p}\right)^{1/p},$$

*where* $c_0 \leqslant 2^{3/4}\sqrt{\pi e} < 5$.

## 2.18 Sparse Sums of Positive Semi-definite Matrices

**Theorem 2.18.1 ([125]).** *Let* $\mathbf{A}_1, \ldots, \mathbf{A}_N$ *be symmetric, positive semidefinite matrices of size* $n \times n$ *and arbitrary rank. For any* $\varepsilon \in (0,1)$*, there is a deterministic algorithm to construct a vector* $\mathbf{y} \in \mathbb{R}^N$ *with* $O(n/\varepsilon^2)$ *nonzero entries such that* $\mathbf{y} \geq \mathbf{0}$ *and*

$$\sum_{i=1}^{N} \mathbf{A}_i \leqslant \sum_{i=1}^{N} y_i \mathbf{A}_i \leqslant (1+\varepsilon) \sum_{i=1}^{N} \mathbf{A}_i.$$

*The algorithm runs in* $O(Nn^3/\varepsilon^2)$ *time. Moreover, the result continues to hold if the input matrices* $\mathbf{A}_1, \ldots, \mathbf{A}_N$ *are Hermitian and positive semi-definite.*

**Theorem 2.18.2 ([125]).** *Let* $\mathbf{A}_1, \ldots, \mathbf{A}_N$ *be symmetric, positive semidefinite matrices of size* $n \times n$ *and let* $\mathbf{y} = (y_1, \ldots, y_N)^T \in \mathbb{R}^N$ *satisfy* $\mathbf{y} \geq \mathbf{0}$ *and* $\sum\limits_{i=1}^{N} y_i = 1$. *For any* $\varepsilon \in (0,1)$*, these exists* $\mathbf{x} \geqslant 0$ *with* $\sum\limits_{i=1}^{N} x_i = 1$ *such that* $\mathbf{x}$ *has* $O(n/\varepsilon)$ *nonzero entries and*

$$(1-\varepsilon)\sum_{i=1}^{N} y_i \mathbf{A}_i \leqslant \sum_{i=1}^{N} x_i \mathbf{A}_i \leqslant (1+\varepsilon) \sum_{i=1}^{N} y_i \mathbf{A}_i.$$

## 2.19 Further Comments

This chapter heavily relies on the work of Tropp [53]. Due to its user-friendly nature, we take so much material from it.

Column subsampling of matrices with orthogonal rows is treated in [111]. Exchangeable pairs for sums of dependent random matrices is studied by Mackey, Jordan, Chen, Farrell, Tropp [126]. Learning with integral operators [127, 128] is relevant. Element-wise matrix sparsification by Drineas [129] is interesting. See [130, Page 15] for some matrix concentration inequalities.

http://www.springer.com/978-1-4614-4543-2

Cognitive Networked Sensing and Big Data
Qiu, R.; Wicks, M.
2014, XXIII, 614 p., Hardcover
ISBN: 978-1-4614-4543-2