Matrix Concentration: Tools

SDS 391P.6, Spring 2026

Pratik Patil

These notes are a work in progress and are provided as-is for instructional purposes only. They are not (yet) at the level of a scholarly document. In particular, the notes draw from various sources and do not (yet) have sufficient references to the original sources. Additionally, almost surely the notes have errors and they are only probably approximately correct. The notes will be updated regularly as the course progresses. Last updated: 2026-03-15.

1 Motivation

So far, our concentration results have been *scalar*: they control random variables such as

$$X_1 + \cdots + X_n, \qquad f(X_1, \ldots, X_n), \qquad \lambda_{\max}(A(X_1, \ldots, X_n)),$$

through tail bounds of the form

$$\mathbb{P}\{Z - \mathbb{E}Z \geqslant t\} \leqslant \cdots .$$

In many problems of modern statistics, machine learning, and random matrix theory, however, the object of interest is itself a *random matrix*. A basic example that we have seen is the sample covariance matrix

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top,$$

which is a random symmetric matrix. In this setting, one often wants to control the deviation of the whole matrix from its mean, for example through quantities like

$$\lambda_{\max}(\widehat{\Sigma} - \mathbb{E}\widehat{\Sigma}) \qquad \text{or} \qquad \|\widehat{\Sigma} - \mathbb{E}\widehat{\Sigma}\|.$$

This is conceptually different from controlling the scalar random variable $\lambda_{\max}(\widehat{\Sigma})$ around its own expectation. Indeed,

$$\mathbb{P}\{\lambda_{\max}(\widehat{\Sigma}) - \mathbb{E}\lambda_{\max}(\widehat{\Sigma}) \geqslant t\}$$

is a concentration statement about one scalar statistic of the matrix, whereas

$$\mathbb{P}\{\lambda_{\max}(\widehat{\Sigma} - \mathbb{E}\widehat{\Sigma}) \geqslant t\}$$

is a matrix deviation bound: it says the random matrix itself stays close to its mean in spectral order.

Why is this stronger point of view useful? Because matrix deviation bounds immediately imply information about:

- all eigenvalues, via Weyl-type perturbation bounds;
- invariant subspaces and eigenvectors, via Davis–Kahan-type perturbation theory;
- all linear functionals of the matrix, since

$$|u^\top(\widehat{\Sigma} - \mathbb{E}\widehat{\Sigma})v| \leqslant \|\widehat{\Sigma} - \mathbb{E}\widehat{\Sigma}\| \cdot \|u\|_2 \|v\|_2.$$

In the scalar setting, the Laplace transform method was one of our main tools: if we could control

$$\psi_X(\theta) = \log \mathbb{E} e^{\theta(X - \mathbb{E}X)},$$

then Chernoff bounds gave exponential tails. Today we begin developing a *matrix* version of this method.

The main difficulty that we will encounter is noncommutativity. For scalar independent random variables, the MGF of a sum factorizes:

$$\mathbb{E} e^{\theta(X_1 + \cdots + X_n)} = \prod_{i=1}^{n} \mathbb{E} e^{\theta X_i}.$$

For matrices, the analogous identity is false in general because

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

unless A and B commute. So the scalar proof does not transfer directly.

The good news is that there is still a beautiful replacement. Instead of exact additivity of scalar cumulant generating functions, one gets a *subadditivity principle for matrix CGFs after taking trace exponentials*. Combined with a matrix version of Chernoff's method, this leads to a general "master theorem" for matrix concentration. This lecture develops that framework.

Throughout this lecture, we work over the reals and with *symmetric* matrices. (Everything extends verbatim to complex Hermitian matrices by replacing transpose by adjoint.)

2 The independent sum model for matrices

The basic model for matrix concentration is exactly the one that worked so well for scalars: an *independent sum*.

Let $X_1, \ldots, X_n$ be independent random symmetric matrices in $\mathbb{S}^d$, and define

$$Y := \sum_{i=1}^{n} X_i.$$

We will seek tail and expectation bounds for

$$\lambda_{\max}(Y - \mathbb{E}Y), \qquad \lambda_{\min}(Y - \mathbb{E}Y), \qquad \|Y - \mathbb{E}Y\|.$$

This model is flexible enough to cover many of the random matrices that arise in practice.

Example: sample covariance. If $x_1, \ldots, x_n \in \mathbb{R}^d$ are i.i.d. copies of a centered random vector x, then

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top$$

is an independent sum of random positive semidefinite matrices. Since

$$\mathbb{E}\widehat{\Sigma} = \mathbb{E}(xx^\top),$$

matrix concentration bounds for $\widehat{\Sigma} - \mathbb{E}\widehat{\Sigma}$ give direct control on the quality of covariance estimation.

Example: graph Laplacians. Many random graph matrices can also be written as sums of independent matrix contributions, one per edge. Matrix concentration then yields spectral control of random graphs.

Example: rectangular matrices. Even though we focus on symmetric matrices in this lecture, rectangular random matrices can often be embedded into symmetric block matrices. We will return to this in the next lecture when we discuss matrix Bernstein and covariance estimation in more detail.

The point is that the independent sum model is the right analog of the scalar sum model, and it is broad enough to encompass many applications without yet forcing us into the full complexity of random matrix theory.

3 Matrix calculus for symmetric matrices

To imitate the scalar Laplace method, we need to make sense of expressions such as

$$e^{\theta X}, \qquad \log \mathbb{E} e^{\theta X}, \qquad \operatorname{tr} e^{\theta X}.$$

So before discussing concentration, we need a small amount of matrix calculus.

3.1 Functional calculus

Let $A \in \mathbb{S}^d$ be symmetric. By the spectral theorem, there exists an orthogonal matrix U and a diagonal matrix

$$\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_d)$$

such that

$$A = U \Lambda U^\top.$$

Equivalently,

$$A = \sum_{i=1}^d \lambda_i u_i u_i^\top,$$

where $u_1, \dots, u_d$ form an orthonormal eigenbasis.

If $f : \mathbb{R} \to \mathbb{R}$ is any function, we define

$$f(A) := U \, f(\Lambda) \, U^\top = \sum_{i=1}^d f(\lambda_i) \, u_i u_i^\top,$$

where

$$f(\Lambda) = \operatorname{diag}(f(\lambda_1), \dots, f(\lambda_d)).$$

So the rule is simple: apply the scalar function to the eigenvalues, and keep the eigenvectors unchanged.

This agrees with familiar operations:

$$A^2 = \sum_{i=1}^d \lambda_i^2 u_i u_i^\top, \qquad A^{-1} = \sum_{i=1}^d \lambda_i^{-1} u_i u_i^\top \quad \text{(if } A \text{ is invertible)},$$

and likewise for square roots, exponentials, logarithms, and so on.

3.2 Matrix exponential and logarithm

The two matrix functions that matter most for concentration are the exponential and logarithm.

If $A \in \mathbb{S}^d$, define

$$e^A := \sum_{i=1}^{d} e^{\lambda_i} u_i u_i^\top.$$

Equivalently, this agrees with the power series

$$e^A = I + A + \frac{A^2}{2!} + \frac{A^3}{3!} + \cdots.$$

The matrix e^A is always positive definite.

If $B \in \mathbb{S}^d$ is positive definite, then all its eigenvalues are strictly positive, and we may define

$$\log B := \sum_{i=1}^{d} (\log \mu_i)\, v_i v_i^\top$$

when

$$B = \sum_{i=1}^{d} \mu_i v_i v_i^\top.$$

Thus $\log B$ is the inverse operation to e^A at the spectral level.

3.3 Loewner order

To compare symmetric matrices, we use the Loewner order.

If $A, B \in \mathbb{S}^d$, we write

$$A \succeq B \qquad \text{if and only if} \qquad A - B \succeq 0,$$

where $M \succeq 0$ means that M is positive semidefinite:

$$u^\top M u \geqslant 0 \qquad \text{for all } u \in \mathbb{R}^d.$$

This is only a partial order, not a total one: two matrices need not be comparable.

A few basic facts are worth keeping in mind.

Eigenvalue monotonicity. If $A \preceq B$, then

$$\lambda_{\max}(A) \leqslant \lambda_{\max}(B), \qquad \lambda_{\min}(A) \leqslant \lambda_{\min}(B).$$

More generally, all ordered eigenvalues are monotone under the Loewner order.

Trace monotonicity under increasing functions. If $A \preceq B$ and $f : \mathbb{R} \to \mathbb{R}$ is increasing, then

$$\operatorname{tr} f(A) \leqslant \operatorname{tr} f(B).$$

This follows because the eigenvalues of A are coordinatewise bounded by those of B.

Congruence preserves order. If $A \preceq B$ and M is any matrix, then

$$MAM^\top \preceq MBM^\top.$$

Indeed,

$$u^\top(MAM^\top)u = (M^\top u)^\top A(M^\top u) \leqslant (M^\top u)^\top B(M^\top u) = u^\top(MBM^\top)u.$$

A warning about matrix monotonicity. For scalars, if f is increasing then $a \leqslant b$ implies $f(a) \leqslant f(b)$. For matrices, this *fails* in general:

$$A \preceq B \implies f(A) \preceq f(B)$$

for arbitrary increasing f. There is a special class of *matrix monotone* functions for which this is true, but not every increasing function is matrix monotone. This is one of the first places where noncommutativity causes real trouble.

4 Matrix MGFs and CGFs

Now let X be a random symmetric matrix in $\mathbb{S}^d$.

We define its matrix moment generating function by

$$\mathbb{M}_X(\theta) := \mathbb{E}e^{\theta X}, \qquad \theta \in \mathbb{R},$$

and its matrix cumulant generating function by

$$\Xi_X(\theta) := \log \mathbb{E}e^{\theta X} = \log \mathbb{M}_X(\theta).$$

These are symmetric matrices.

A few comments are in order.

First, $\mathbb{M}_X(\theta)$ is positive definite for every θ for which it is finite, because $e^{\theta X}$ is positive definite and expectations preserve positive definiteness.

Second, just as in the scalar setting, the MGF formally packages moments:

$$\mathbb{M}_X(\theta) = I + \theta\mathbb{E}X + \frac{\theta^2}{2}\mathbb{E}[X^2] + \cdots.$$

Likewise, the matrix CGF has a second-order term that plays the role of a matrix variance proxy. We will not need the full expansion, but the analogy is helpful.

Third, the matrix CGF is *not* additive in the naive scalar sense: for independent X and Y, one does *not* generally have

$$\Xi_{X+Y}(\theta) = \Xi_X(\theta) + \Xi_Y(\theta).$$

This is the main structural obstruction in matrix concentration, and overcoming it will be one of the central achievements of this lecture.

5 The matrix Laplace transform method

We now prove the matrix analog of Chernoff's method. It bounds the upper tail of the maximum eigenvalue of a random symmetric matrix in terms of the trace of a matrix exponential.

Proposition 5.1 (Matrix Laplace transform method)**.** *Let $Y \in \mathbb{S}^d$ be a random symmetric matrix. Then for every $t \in \mathbb{R}$,*

$$\mathbb{P}\{\lambda_{\max}(Y) \geqslant t\} \leqslant \inf_{\theta>0} e^{-\theta t}\, \mathbb{E}\operatorname{tr} e^{\theta Y}.$$

Equivalently,

$$\mathbb{P}\{\lambda_{\max}(Y) \geqslant t\} \leqslant \inf_{\theta>0} \exp\Big(-\theta t + \log \mathbb{E}\operatorname{tr} e^{\theta Y}\Big).$$

Proof. Fix $\theta > 0$. By Markov's inequality,

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

Now use the spectral mapping theorem:

$$e^{\theta\lambda_{\max}(Y)} = \lambda_{\max}(e^{\theta Y}).$$

Since $e^{\theta Y}$ is positive definite, all its eigenvalues are positive, so its largest eigenvalue is bounded by its trace:

$$\lambda_{\max}(e^{\theta Y}) \leqslant \operatorname{tr} e^{\theta Y}.$$

Therefore

$$\mathbb{P}\{\lambda_{\max}(Y) \geqslant t\} \leqslant e^{-\theta t}\, \mathbb{E}\operatorname{tr} e^{\theta Y}.$$

Taking the infimum over $\theta > 0$ completes the proof. □

This is the exact matrix analog of the scalar Chernoff method, except for one important change: the scalar exponential moment $\mathbb{E}e^{\theta Y}$ is replaced by $\mathbb{E}\operatorname{tr} e^{\theta Y}$. That trace is not an artifact of the proof; it is the natural scalar quantity that controls the maximal eigenvalue.

5.1 Expectation bound

The same method also yields a bound on the expectation of the top eigenvalue.

Corollary 5.2 (Expectation form of matrix Laplace)**.** *Let $Y \in \mathbb{S}^d$ be a random symmetric matrix. Then*

$$\mathbb{E}\lambda_{\max}(Y) \leqslant \inf_{\theta>0} \frac{1}{\theta} \log \mathbb{E}\operatorname{tr} e^{\theta Y}.$$

Proof. For every $\theta > 0$, Jensen's inequality for the convex function $x \mapsto e^{\theta x}$ gives

$$e^{\theta\mathbb{E}\lambda_{\max}(Y)} \leqslant \mathbb{E} e^{\theta\lambda_{\max}(Y)} = \mathbb{E}\lambda_{\max}(e^{\theta Y}) \leqslant \mathbb{E}\operatorname{tr} e^{\theta Y}.$$

Taking logarithms and dividing by θ gives the claim. □

5.2 Lower tail

Since

$$\lambda_{\min}(Y) = -\lambda_{\max}(-Y),$$

the same argument applied to $-Y$ yields

$$\mathbb{P}\{\lambda_{\min}(Y) \leqslant t\} = \mathbb{P}\{\lambda_{\max}(-Y) \geqslant -t\} \leqslant \inf_{\theta<0} e^{-\theta t}\, \mathbb{E}\,\mathrm{tr}\, e^{\theta Y}.$$

So upper and lower spectral tails are handled symmetrically.

6 Why naive MGF factorization fails

At this point, the scalar analogy suggests the next step. If

$$Y = \sum_{i=1}^{n} X_i$$

with independent summands, then in the scalar setting we would write

$$\mathbb{E}e^{\theta \sum_i X_i} = \prod_i \mathbb{E}e^{\theta X_i},$$

or equivalently

$$\psi_{\sum_i X_i}(\theta) = \sum_i \psi_{X_i}(\theta).$$

For matrices, the problem is immediate:

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

in general, because matrix multiplication is noncommutative. Thus the scalar MGF factorization breaks at its first step.

One might hope that the trace fixes the issue, but even that is not true in any simple exact way. There is a famous inequality of Golden–Thompson:

$$\mathrm{tr}\, e^{A+B} \leqslant \mathrm{tr}(e^A e^B),$$

which is extremely useful for two matrices. However, there is no comparable exact product formula for arbitrary sums of many independent random matrices. So some deeper idea is required.

The right replacement turns out to be this: although matrix MGFs do not factorize, matrix CGFs are *subadditive after applying* $\mathrm{tr}\exp$. This is a subtle and very powerful fact, and it is where Lieb's theorem enters.

7 Lieb's concavity theorem

We now state the matrix-analytic result that makes the whole theory work.

Theorem 7.1 (Lieb's concavity theorem)**.** *Fix $H \in \mathbb{S}^d$. Then the map*

$$A \mapsto \operatorname{tr} \exp\bigl(H + \log A\bigr)$$

is concave on the positive definite cone

$$\mathbb{S}^d_{++} := \{A \in \mathbb{S}^d : \ A \succ 0\}.$$

This theorem is deep. We will not prove it here. For us, its role is similar to the role played by Hölder/Jensen in the scalar Laplace method: it is the structural inequality that compensates for the failure of commutativity.

The form in which we will use it is the following corollary.

Corollary 7.2 (Lieb + Jensen)**.** *Let $H \in \mathbb{S}^d$ be deterministic, and let $X \in \mathbb{S}^d$ be random. Then*

$$\mathbb{E} \operatorname{tr} \exp(H + X) \leqslant \operatorname{tr} \exp\bigl(H + \log \mathbb{E} e^X\bigr).$$

Proof. Set

$$F(A) := \operatorname{tr} \exp(H + \log A).$$

By Theorem 7.1, F is concave on $\mathbb{S}^d_{++}$. Apply Jensen's inequality to the positive definite random matrix e^X:

$$\mathbb{E} F(e^X) \leqslant F(\mathbb{E} e^X).$$

Expanding the definition of F,

$$\mathbb{E} \operatorname{tr} \exp(H + \log e^X) \leqslant \operatorname{tr} \exp(H + \log \mathbb{E} e^X).$$

Since $\log e^X = X$, this becomes

$$\mathbb{E} \operatorname{tr} \exp(H + X) \leqslant \operatorname{tr} \exp\bigl(H + \log \mathbb{E} e^X\bigr).$$

□

This corollary is the key iterative device behind the subadditivity theorem below.

8 Subadditivity of matrix CGFs

We can now state and prove the matrix analog of scalar CGF additivity.

Theorem 8.1 (Subadditivity of matrix CGFs)**.** *Let $X_1, \dots, X_n$ be independent random symmetric matrices in $\mathbb{S}^d$. Then for every $\theta \in \mathbb{R}$,*

$$\mathbb{E} \operatorname{tr} \exp\left(\theta \sum_{i=1}^n X_i\right) \leqslant \operatorname{tr} \exp\left(\sum_{i=1}^n \Xi_{X_i}(\theta)\right),$$

where

$$\Xi_{X_i}(\theta) := \log \mathbb{E} e^{\theta X_i}.$$

Equivalently,

$$\mathbb{E} \operatorname{tr} \exp\left(\theta \sum_{i=1}^n X_i\right) \leqslant \operatorname{tr} \exp\left(\sum_{i=1}^n \log \mathbb{E} e^{\theta X_i}\right).$$

Proof. We prove the result by iterating Corollary 7.2.

Write

$$Y_n := \theta \sum_{i=1}^{n} X_i.$$

Condition on $X_1, \dots, X_{n-1}$ and view

$$H := \theta \sum_{i=1}^{n-1} X_i$$

as fixed. Applying Corollary 7.2 to the last summand θX_n,

$$\mathbb{E}_{X_n} \operatorname{tr} \exp(H + \theta X_n) \leqslant \operatorname{tr} \exp\Big(H + \log \mathbb{E} e^{\theta X_n}\Big).$$

Now take expectation over $X_1, \dots, X_{n-1}$:

$$\mathbb{E} \operatorname{tr} \exp\left(\theta \sum_{i=1}^{n} X_i\right) \leqslant \mathbb{E} \operatorname{tr} \exp\left(\theta \sum_{i=1}^{n-1} X_i + \log \mathbb{E} e^{\theta X_n}\right).$$

The last term is again of the same form: a random sum plus a deterministic matrix. Now repeat the argument with X_{n-1}, then X_{n-2}, and so on. After n steps, all randomness is removed, and we obtain

$$\mathbb{E} \operatorname{tr} \exp\left(\theta \sum_{i=1}^{n} X_i\right) \leqslant \operatorname{tr} \exp\left(\sum_{i=1}^{n} \log \mathbb{E} e^{\theta X_i}\right).$$

This is the desired inequality. □

This theorem is the exact replacement for scalar CGF additivity that we needed. It is not an equality, but it is strong enough for concentration purposes.

9 The master theorem for matrix concentration

We now combine the matrix Laplace transform method with subadditivity of matrix CGFs.

Theorem 9.1 (Master tail bound for independent sums)**.** *Let $X_1, \dots, X_n$ be independent random symmetric matrices in $\mathbb{S}^d$, and set*

$$Y := \sum_{i=1}^{n} X_i.$$

Then for every $t \in \mathbb{R}$,

$$\mathbb{P}\{\lambda_{\max}(Y) \geqslant t\} \leqslant \inf_{\theta > 0} e^{-\theta t} \operatorname{tr} \exp\left(\sum_{i=1}^{n} \Xi_{X_i}(\theta)\right),$$

where

$$\Xi_{X_i}(\theta) = \log \mathbb{E} e^{\theta X_i}.$$

Equivalently,

$$\mathbb{P}\{\lambda_{\max}(Y) \geqslant t\} \leqslant \inf_{\theta > 0} \exp\left(-\theta t + \log \operatorname{tr} \exp\left(\sum_{i=1}^{n} \Xi_{X_i}(\theta)\right)\right).$$

Likewise,

$$\mathbb{P}\{\lambda_{\min}(Y) \leqslant t\} \leqslant \inf_{\theta < 0} e^{-\theta t} \operatorname{tr} \exp\left(\sum_{i=1}^{n} \Xi_{X_i}(\theta)\right).$$

Proof. Apply Proposition 5.1 to Y:

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

Now use Theorem 8.1:

$$\mathbb{E}\,\mathrm{tr}\, e^{\theta Y} = \mathbb{E}\,\mathrm{tr}\exp\left(\theta \sum_{i=1}^{n} X_i\right) \leqslant \mathrm{tr}\exp\left(\sum_{i=1}^{n} \Xi_{X_i}(\theta)\right).$$

Substituting this into the Laplace bound gives the upper-tail inequality. The lower-tail bound follows by applying the same argument to $-Y$, or equivalently by using the $\theta < 0$ form of the Laplace method. □

We also obtain the expectation form immediately.

Corollary 9.2 (Master expectation bound)**.** *Under the assumptions of Theorem 9.1,*

$$\mathbb{E}\lambda_{\max}(Y) \leqslant \inf_{\theta>0} \frac{1}{\theta} \log \mathrm{tr}\exp\left(\sum_{i=1}^{n} \Xi_{X_i}(\theta)\right).$$

Likewise,

$$\mathbb{E}\lambda_{\min}(Y) \geqslant \sup_{\theta<0} \frac{1}{\theta} \log \mathrm{tr}\exp\left(\sum_{i=1}^{n} \Xi_{X_i}(\theta)\right).$$

Proof. Combine Corollary 5.2 with Theorem 8.1. The lower bound follows by applying the upper bound to $-Y$. □

The master theorem is quite powerful. Once we can upper bound each individual matrix CGF $\Xi_{X_i}(\theta)$, we get a concentration inequality for the whole sum. This is completely parallel to the scalar story:

$$\text{bound each scalar CGF } \psi_{X_i} \quad \Longrightarrow \quad \text{control } \psi_{\sum X_i} \quad \Longrightarrow \quad \text{Chernoff tail bound.}$$

The only difference is that in the matrix world, additivity of CGFs is replaced by the subadditivity theorem above.

10 Look ahead

At this point, we have reduced matrix concentration to a deterministic-looking problem: control the matrix CGFs of the summands.

The master theorem tells us that if we can prove a Loewner-order bound of the form

$$\Xi_{X_i}(\theta) \preceq G_i(\theta)$$

for some explicit matrices $G_i(\theta)$, then

$$\sum_{i=1}^{n} \Xi_{X_i}(\theta) \preceq \sum_{i=1}^{n} G_i(\theta),$$

and trace monotonicity yields

$$\operatorname{tr}\exp\left(\sum_{i=1}^{n}\Xi_{X_i}(\theta)\right) \leqslant \operatorname{tr}\exp\left(\sum_{i=1}^{n}G_i(\theta)\right).$$

So the problem becomes:

$$\textit{Find tractable upper bounds for } \log \mathbb{E}e^{\theta X_i}.$$

This is the matrix analog of Hoeffding's lemma and Bernstein's MGF bound from the scalar setting.

In the next lecture, we will do exactly that. We will derive:

- matrix Hoeffding inequalities for bounded summands;
- matrix Bernstein inequalities for bounded or sub-exponential-type summands;
- applications to covariance estimation beyond the sub-Gaussian setting.

Source material

Parts of this lecture are based on references: Tropp (2023); Vershynin (2018), in addition to the author's accumulated experience working on related topics.

References

Tropp, J. A. (2023). Probability in high dimensions. Caltech CMS Lecture Notes 2021-01.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.