# CS395T: Continuous Algorithms, Part VI
# Matrix analysis and concentration

Kevin Tian

## 1 Scalar concentration

In modern data science, it is often useful to model datasets as being drawn from a high-dimensional distribution. To predict how algorithms behave on datasets, we frequently wish to establish some basic properties about the corresponding distribution: what does a typical dataset look like, and are statistics of the dataset well-behaved, i.e., do they concentrate around their typical values? This prompts the development of *matrix concentration inequalities*, where we ask quantitative questions about the behavior of high-dimensional data. To build up to these tools, we first establish some basic techniques which are useful in the study of how scalar random variables concentrate.

### 1.1 Sub-Gaussian random variables

Let $\{Z_i\}_{i\in[n]} \in \{\pm 1\}^n$ be a collection of independent Rademacher random variables, i.e., let

$$Z_i = \begin{cases} 1 & \text{with probability } \frac{1}{2} \\ -1 & \text{with probability } \frac{1}{2} \end{cases} \text{ for all } i \in [n].$$

For example, $Z_i$ could measure whether the $i^{\text{th}}$ toss of a fair coin came up heads. One basic statistic of $\{Z_i\}_{i\in[n]}$ is its sum $Z = \sum_{i\in[n]} Z_i$: when tossing $n$ coins, what is the typical discrepancy between the number of heads and tails? If we only care about a rough estimate, we can apply Chebyshev's inequality, which is a consequence of the simpler Markov's inequality applied to $|Z - \mathbb{E}Z|^2$.

**Fact 1** (Markov and Chebyshev)**.** *If $Z$ is a nonnegative random variable, $\Pr[Z \geq t] \leq \frac{1}{t}\mathbb{E}Z$ for all $t \geq 0$. If $Z$ is a random variable taking values in $\mathbb{R}$, $\Pr[|Z - \mathbb{E}Z| \geq t] \leq \frac{1}{t^2}\mathrm{Var}[Z]$.*

Clearly, $\mathbb{E}Z = 0$, and further, as $\mathrm{Var}[X_i] = 1$ for all $i \in [n]$, we have $\mathrm{Var}[Z] = n$ by independence. Therefore, Fact 1 shows that for any $\delta \in (0, 1)$, we have $|Z| = \sqrt{n/\delta}$ with probability $1-\delta$. However, we should be able to say something better: as $n \to \infty$, the central limit theorem says that $\frac{1}{n}Z$ is approximately distributed as $\mathcal{N}(0, \frac{1}{n})$, and the tail behavior of Gaussians is significantly better than what Chebyshev predicts. Specifically, assuming $Z \sim \mathcal{N}(0, n)$ (i.e., accepting the central limit theorem approximation as fact), standard tail bounds on Gaussian random variables show $|Z| = O(\sqrt{n \log(1/\delta)})$ except with probability $\delta$. This is a significantly improved bound in the regime $\delta \to 0$, and as we will see it continues to hold for true for actual Rademacher sums.

The key definition which helps in establishing this is the following.

**Definition 1** (Sub-Gaussian)**.** *We say real-valued random variable $Z$ is $\sigma^2$-*sub-Gaussian *if*

$$\mathbb{E}\exp\left(\lambda\left(Z - \mathbb{E}Z\right)\right) \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right) \text{ for all } \lambda \in \mathbb{R}. \tag{1}$$

When $Z$ is centered (i.e., $\mathbb{E}Z = 0$), the left-hand side of (1) is simply the moment-generating function (MGF) of $Z$. Direct computation of the MGF shows that when $Z \sim \mathcal{N}(\mu, \sigma^2)$ for any $\mu \in \mathbb{R}$, (1) holds with equality, reflecting the intuition that $\sigma^2$-sub-Gaussianity of $Z$ serves as a proxy for the statement "$Z$ behaves like a draw $\sim \mathcal{N}(0, \sigma^2)$." Moreover, sums of independent sub-Gaussian random variables are also sub-Gaussian, by expanding (1) and applying independence.

**Fact 2.** *If $\{Z_i\}_{i\in[n]}$ are real-valued random variables such that $Z_i$ is $\sigma_i^2$-sub-Gaussian for all $i \in [n]$, then the random variable $\sum_{i\in[n]} Z_i$ is $\sum_{i\in[n]} \sigma_i^2$-sub-Gaussian.*

Our Rademacher example is thus also sub-Gaussian, as captured by the following general statement.

**Lemma 1.** *Let $Z$ be a random variable supported in $[a, b]$. Then $Z$ is $(b-a)^2$-sub-Gaussian.*[1]

*Proof.* Let $X = \frac{1}{b-a}(Z - \mathbb{E}Z)$ so that $X \in [-1, 1]$, $\mathbb{E}X = 0$.[2] We claim $X$ is 1-sub-Gaussian, which yields the claim because (1) holds for all $\lambda \in \mathbb{R}$, so we can scale $\lambda$ by $b-a$ so the definitions agree:

$$\mathbb{E}\exp\left(\lambda(Z - \mathbb{E}Z)\right) = \mathbb{E}\exp\left(\lambda(b-a)X\right) \leq \exp\left(\frac{\lambda^2(b-a)^2}{2}\right).$$

Next, we claim that for all $\lambda \in \mathbb{R}$, $\mathbb{E}\exp(\lambda X) \leq \mathbb{E}\exp(\lambda\sigma)$, where $\sigma \in \{\pm 1\}$ is a Rademacher random variable. To see this, for each conditional realization $X = t$ with $t \in [-1, 1]$, consider "spreading" $t$ to the endpoints $\{\pm 1\}$ in an expectation-preserving way, i.e. let $(Y \mid X = t) \in \{\pm 1\}$ and $\mathbb{E}[Y \mid X = t] = t$. Because $f(t) = \exp(\lambda t)$ is convex, Jensen's inequality shows

$$\mathbb{E}\exp(\lambda X) \leq \mathbb{E}\left[\mathbb{E}\left[\exp(\lambda Y) \mid X = t\right]\right] = \mathbb{E}\left[\exp(\lambda Y)\right].$$

Also, note that $\mathbb{E}Y = 0$, so $Y$ is distributionally equivalent to $\sigma$. Finally, we have the claim due to the straightforward calculation (which follows by Taylor expansion):

$$\mathbb{E}[\exp(\lambda\sigma)] = \frac{1}{2}\exp(\lambda) + \frac{1}{2}\exp(-\lambda) \leq \exp\left(\frac{\lambda^2}{2}\right).$$

□

By combining Fact 2 and Lemma 1, we have established that our earlier Rademacher sum is $O(n)$-sub-Gaussian. This yields our earlier claimed concentration bound via the following observation.

**Theorem 1** (Hoeffding's inequality)**.** *Let $Z$ be $\sigma^2$-sub-Gaussian. Then for all $t \geq 0$,*

$$\Pr\left[|Z - \mathbb{E}Z| \geq t\right] \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right).$$

*Proof.* By Markov's inequality (Fact 1) and (1), for any $\lambda \geq 0$,

$$\Pr\left[Z - \mathbb{E}Z \geq t\right] = \Pr\left[\exp\left(\lambda(Z - \mathbb{E}Z)\right) \geq \exp(\lambda t)\right] \leq \exp\left(-\lambda t + \frac{\lambda^2\sigma^2}{2}\right). \tag{2}$$

The proof follows by optimizing in $\lambda$, and observing that the argument is symmetric in $Z, -Z$. □

**Remark 1.** *The proof technique in Theorem 1, also known as the MGF method, is intimately connected to our discussion of convex conjugates in Part III. Letting $\psi(\lambda) := \log \mathbb{E}\exp(\lambda(Z - \mathbb{E}(Z)))$ be the logarithm of the left-hand side in (1), the Markov argument in (2) shows that*

$$\Pr[Z - \mathbb{E}Z \geq t] \leq \exp\left(-\left(\sup_{\lambda \in \mathbb{R}} \lambda t - \psi(\lambda)\right)\right) = \exp\left(-\psi^*(t)\right),$$

*so it suffices to bound $\psi^*$, the* Cramér transform *of $Z$.*[3] *Theorem 1 is a simple consequence of order-preserving properties of conjugation,*[4] *and that $\psi^*(t) = \frac{t^2}{2\sigma^2}$ when $\psi(\lambda) = \frac{\lambda^2\sigma^2}{2}$. Similar arguments appear in many other proofs; for a more complete explanation, see Section 2.2 of [BLM13].*

Indeed, Theorem 1 shows that for any $\delta \in (0, 1)$, an $O(n)$-sub-Gaussian random variable does not deviate from its mean by $\omega(\sqrt{n\log(1/\delta)})$ except with probability $\delta$, as claimed before. We conclude by mentioning some equivalent characterizations of sub-Gaussian random variables which are often useful in applications, deferring a proof to Proposition 2.5.2 of [Ver18].[5]

[1] This can be improved to show that $Z$ is in fact $\frac{1}{4}(b-a)^2$-sub-Gaussian; for a (more complicated, and thus omitted) proof, we refer the reader to the Wikipedia page for "Hoeffding's lemma."

[2] The lossiness in this proof is because $X$ is actually supported in $[-c, 1-c]$ for some $c \in [0, 1]$.

[3] Here, we used that $\psi(\lambda) \geq 0$ for all $\lambda$ by convexity of log, so the maximal $\lambda$ in the expression $\sup_{\lambda \in \mathbb{R}} \lambda t - \psi(\lambda)$ will never be negative and hence we can extend $\lambda \geq 0$ to $\lambda \in \mathbb{R}$ in the expression.

[4] That is, if $\psi \leq \varphi$ pointwise then $\psi^* \geq \varphi^*$ pointwise, by the definition of conjugates.

[5] A proof of the moment bound can be found in (4).

**Proposition 1.** *The following properties are equivalent, for constants $C_1$, $C_2$, and $C_3$.*

1. Sub-Gaussianity*: $Z$ is $\sigma^2$-sub-Gaussian.*
2. Moment bound*: For all $p \geq 1$, $\mathbb{E}|Z|^p \leq (C_1\sigma\sqrt{p})^p$.*
3. Tail bound*: For all $t \geq 0$, $\Pr[|Z| \geq t] \leq 2\exp(-(\frac{t}{C_2\sigma})^2)$.*
4. MGF of square bound*: $\mathbb{E}[\exp(\frac{Z}{C_3\sigma})^2] \leq 2$.*

## 1.2 Sub-gamma random variables

In many applications, it is helpful to consider a truncated variant of sub-Gaussianity where (1) only holds for a range of $\lambda$. For example, when bounding empirical second moments we often require concentration behavior of the square of a Gaussian, which is not sub-Gaussian, as we can check by computing the MGF. Fortunately, it does satisfy a weaker definition which we now introduce.[6]

**Definition 2** (Sub-gamma)**.** *We say $Z$ is $(\sigma^2, c)$-*sub-gamma *if (1) holds for all $|\lambda| \leq \frac{1}{c}$.*

The symmetrized gamma distribution with parameters $a, b$ is $(O(ab^2), O(b))$-sub-gamma. We next give two more canonical examples of sub-gamma random variables with simple interpretations.

**Lemma 2.** *If $Z$ is supported in $[-c, c]$ with $\mathbb{E}Z = 0$ and $\mathbb{E}Z^2 = \sigma^2$, $Z$ is $(2\sigma^2, 2c)$-sub-gamma.*

*Proof.* First, note that for any $p \geq 2$, $\mathbb{E}[Z^p] \leq \mathbb{E}[Z^2c^{p-2}] = \sigma^2c^{p-2}$. We therefore have by a Taylor expansion that for $|\lambda| \leq \frac{1}{2c}$, using $\mathbb{E}Z = 0$ and Stirling's approximation $p! \geq (\frac{p}{e})^p$,

$$\mathbb{E}[\exp(\lambda Z)] = 1 + \mathbb{E}\left[\sum_{p=2}^{\infty} \frac{\lambda^p Z^p}{p!}\right] \leq 1 + \sum_{p=2}^{\infty} \frac{\lambda^p\sigma^2c^{p-2}}{p!} \leq 1 + \sum_{p=2}^{\infty} \frac{\lambda^2\sigma^2}{2^{p-1}} \leq \exp\left(\lambda^2\sigma^2\right). \tag{3}$$

□

Lemma 2 gives a tighter characterization of the MGF of $Z$ than Lemma 1, which only offers a sub-Gaussianity bound of $O(c^2)$, because $\sigma^2$ can be significantly smaller than $c^2$ if $Z$ is tightly centered. The tradeoff is that we can only control the MGF of $Z$ for bounded $|\lambda| \leq \frac{1}{2c}$.

**Lemma 3.** *Let $Z$ be $\sigma^2$-sub-Gaussian. Then $(Z - \mathbb{E}Z)^2$ is $(80\sigma^4, 40\sigma^2)$-sub-gamma.*

*Proof.* Let $X := Z - \mathbb{E}Z$. We first observe that by sub-Gaussianity of $Z$ and Theorem 1,

$$\begin{aligned}\mathbb{E}|X|^p &= \int_0^\infty \Pr[|X|^p \geq t]\mathrm{d}t = \int_0^\infty (pt^{p-1})\Pr[|X| \geq t]\mathrm{d}t \\ &\leq 2p\int_0^\infty t^{p-1}\exp\left(-\frac{t^2}{2\sigma^2}\right)\mathrm{d}t = p\sigma^p\Gamma\left(\frac{p}{2}\right) \leq (2\sigma\sqrt{p})^p,\end{aligned} \tag{4}$$

where $\Gamma$ is the Gamma function. In the first line we did a change of variables, and in the second line we used standard bounds on $\Gamma$. For $|\lambda| \leq \frac{1}{40\sigma^2}$, we therefore have the desired claim from

$$\begin{aligned}\mathbb{E}\exp(\lambda(X^2 - \mathbb{E}X^2)) &= 1 + \sum_{p=2}^{\infty}\frac{1}{p!}\mathbb{E}[\lambda^p(X^2 - \mathbb{E}X^2)^p] \leq 1 + \sum_{p=2}^{\infty}\frac{1}{p!}\mathbb{E}[\lambda^p X^{2p}] \\ &\leq 1 + \sum_{p=2}^{\infty}\frac{(8\lambda\sigma^2 p)^p}{p!} \leq 1 + \sum_{p=2}^{\infty}(8e\lambda\sigma^2)^p \leq \exp\left(40\lambda^2\sigma^4\right).\end{aligned}$$

The first inequality used that for any nonnegative random variable $Y$, it is the case that[7] $\mathbb{E}[(Y - \mathbb{E}Y)^p] \leq \mathbb{E}[Y^p]$ for any $p \geq 0$, the second used (4), and the third used Stirling's approximation. □

We note that $(c^2, c)$-sub-gamma random variables are sometimes called $c$-sub-exponential in the literature, because the exponential distribution with density $\propto \exp(-\frac{\cdot}{c})$ has this property. We next observe that Facts 2 and Theorem 1 have straightforward extensions to the sub-gamma setting.

[6] Different sources have slightly different definitions (e.g., [BLM13, Duc23]), but they are identical up to constants.

[7] This follows from taking expectations over the inequality $y^p - (y-\mu)^p \geq y - \mu$, which holds for all $y, \mu \geq 0$ and $p \geq 2$, since the expectation of the right-hand side is 0 for $\mu = \mathbb{E}Y$.

**Fact 3.** *If $\{Z_i\}_{i\in[n]}$ are real-valued random variables such that $Z_i$ is $(\sigma_i^2, c_i)$-sub-gamma for all $i \in [n]$, then the random variable $\sum_{i\in[n]} Z_i$ is $(\sum_{i\in[n]} \sigma_i^2, \max_{i\in[n]} c_i)$-sub-gamma.*

**Theorem 2** (Bernstein's inequality)**.** *Let $Z$ be $(\sigma^2, c)$-sub-gamma. Then for all $t \geq 0$,*

$$\Pr[|Z - \mathbb{E}Z| \geq t] \leq 2\exp\left(-\min\left(\frac{t^2}{2\sigma^2}, \frac{t}{2c}\right)\right).$$

*Proof.* If $\frac{t}{\sigma^2} \leq \frac{1}{c}$, we can use the proof of Theorem 1; else, we plug $\lambda = \frac{1}{c}$ in (2). □

As a first application of Theorem 2, consider a $\chi^2$ random variable with $d$ degrees of freedom, i.e., $Z = \sum_{i\in[d]} \boldsymbol{\gamma}_i^2$ where we draw i.i.d. $\{\boldsymbol{\gamma}_i\}_{i\in[d]} \sim \mathcal{N}(0,1)$. Combining Lemma 3 and Fact 3 shows $Z$ is $(O(d), O(1))$-sub-gamma, and hence Theorem 2 shows that with probability $\geq 1 - \delta$,

$$|Z - d| \leq O\left(\sqrt{d\log\left(\frac{1}{\delta}\right)} + \log\left(\frac{1}{\delta}\right)\right).$$

This well-known tail bound of [LM00] shows that the squared norms of random Gaussian vectors concentrate extremely tightly around their mean of $d$, with a typically sublinear deviation $\approx \sqrt{d}$.

Theorem 2 shows that the tail behavior of sub-gamma random variables has two phases, with a transition at $t^\star = \frac{\sigma^2}{c}$. Below the critical value $t^\star$, they behave like sub-Gaussians, and above it they deviate from their expectations by $\omega(c\log(\frac{1}{\delta}))$ with probability $\leq \delta$. We conclude with an application of these tools which is very powerful in algorithm design, as we will see in later lectures.

**Corollary 1** (Johnson-Lindenstrauss)**.** *Let $\{\mathbf{g}_i\}_{i\in[k]} \subseteq \mathbb{R}^d$ be i.i.d. draws from $\mathcal{N}(\mathbf{0}_d, \frac{1}{k}\mathbf{I}_d)$, and let $\mathbf{v} \in \mathbb{R}^d$ be fixed. Then for $Z := \sum_{i\in[k]} \langle \mathbf{g}_i, \mathbf{v}\rangle^2 = \|\mathbf{G}\mathbf{v}\|_2^2$, where $\mathbf{G} \in \mathbb{R}^{k\times d}$ has rows $\{\mathbf{g}_i\}_{i\in[k]}$,*[8]

$$\Pr\left[|Z - \|\mathbf{v}\|_2^2| \geq \epsilon \|\mathbf{v}\|_2^2\right] \leq 2\exp\left(-\frac{\epsilon^2 k}{160}\right), \text{ for all } \epsilon \in \left(0, \frac{1}{2}\right).$$

*Proof.* Let $\sigma := k^{-1/2}\|\mathbf{v}\|_2$, and note that each $\langle \mathbf{g}_i, \mathbf{v}\rangle$ is distributed as $\mathcal{N}(0, \frac{1}{k}\|\mathbf{v}\|_2^2)$, so it is zero-mean and $\sigma^2$-sub-Gaussian. Moreover, clearly $\mathbb{E}Z = \|\mathbf{v}\|_2^2$. Hence, combining Lemma 3 and Fact 3 shows $Z$ is $(\frac{80}{k}\|\mathbf{v}\|_2^4, \frac{40}{k}\|\mathbf{v}\|_2^2)$-sub-gamma, and the conclusion follows from Theorem 2. □

Corollary 1 shows that if we choose $k = \Theta(\log(\frac{1}{\delta}) \cdot \epsilon^{-2})$ for an appropriate constant, a random Gaussian projection $\mathbf{G}\mathbf{v}$ preserves the norm of $\mathbf{v}$ up to a $1 \pm \epsilon$ factor except with probability $\delta$. Importantly, $k \ll d$ in many parameter regimes, and a union bound lets us reuse the same "sketching matrix" $\mathbf{G}$ for a set of multiple vectors, which can significantly speed up distance computations. For example, the Johnson-Lindenstrauss lemma is often rephrased as the following consequence of Corollary 1: given $n$ vectors $\{\mathbf{v}_i\}_{i\in[n]} \subset \mathbb{R}^d$, their projections $\{\mathbf{u}_i := \mathbf{G}\mathbf{v}_i\}_{i\in[n]}$ have each pairwise distance $\|\mathbf{u}_i - \mathbf{u}_j\|_2$ within a $1\pm\epsilon$ factor of each $\|\mathbf{v}_i - \mathbf{v}_j\|_2$, when $\mathbf{G} \in \mathbb{R}^{k\times d}$ has $k = \Theta(\log(n)\cdot\epsilon^{-2})$.

## 1.3 Chernoff bounds

Consider again our example of coin flips, but assume instead that we are flipping $n$ biased coins with heads probability $p = \frac{\mu}{n}$ for fixed $\mu$, i.e., as $n \to \infty$ seeing a head is a very rare event. The limiting behavior of $Z$, the number of heads observed, is a Poisson random variable with rate parameter $\mu$, i.e., $\Pr[Z = k] \approx \frac{1}{k!}\mu^k\exp(-\mu)$ for all $k \geq 0$.[9] Observe that

$$\frac{1}{k!}\mu^k\exp(-\mu) \approx \left(\frac{e\mu}{k}\right)^k\exp(-\mu) = \exp(-\Omega(k\log k)), \tag{5}$$

if we treat $\mu$ as constant. This shows that the tail behavior of $Z$ is slightly better than an exponential random variable, but worse than a sub-Gaussian random variable for large $k$.

Now, let us see what Theorems 1 and 2 would predict. Combining Lemma 1 and Fact 2 shows $Z$ is $n$-sub-Gaussian, so Theorem 1 shows that with high probability, $Z \leq \mu + O(\sqrt{n})$. On the other

[8] The constant can be significantly improved via a more precise variant of Lemma 3 when $Z$ is Gaussian.
[9] This is often called the Poisson limit theorem.

hand, since the variance of each coin (treated as a $\{0,1\}$ variable) is $\leq \frac{\mu}{n}$, combining Lemma 2 and Fact 3 shows that $Z$ is $(2\mu, 2)$-sub-gamma, so Theorem 1 predicts that with high probability, $Z \leq O(\max(\mu, 1))$. Interestingly, in small regimes of $\mu$, both bounds are lossier than what we can show using the nonnegativity structure of $Z$; such structure-aware bounds in this setting are known as *Chernoff bounds*. We state a fairly general version of the Chernoff bound, and then interpret it.

**Theorem 3** (Chernoff bound)**.** *Let $\{Z_i\}_{i\in[n]}$ be independent random variables supported in $[0, R]$, let $Z = \sum_{i\in[n]} Z_i$, and let $\mu := \mathbb{E}Z$. Then,*

$$\Pr\left[Z \geq (1+\epsilon)\mu\right] \leq \exp\left(-\frac{\epsilon^2\mu}{(2+\epsilon)R}\right) \text{ for all } \epsilon \geq 0,$$
$$\Pr\left[Z \leq (1-\epsilon)\mu\right] \leq \exp\left(-\frac{\epsilon^2\mu}{2R}\right) \text{ for all } \epsilon \in (0,1).$$

We mention that similarly to Theorem 2, this result has a phase transition which occurs roughly when $\epsilon \approx 1$. Theorem 3 improves our earlier characterizations (by way of Theorems 1 and 2) in this regime, because it allows for very tight control of constants. For example, if $R = 1$, $\epsilon \in (0,1)$ and $\mu \gg \frac{1}{\epsilon^2}$, Theorem 3 shows that $Z$ falls in the range $[(1-\epsilon)\mu, (1+\epsilon)\mu]$ with high probability, whereas our previous characterizations lost constant factors or worse. The following mnemonic is useful for remembering Theorem 3, and captures many of its major applications.

**Corollary 2.** *Let $Z = \sum_{i\in[n]} Z_i$ be a* Poisson binomial *random variable, where each $Z_i \in \{0,1\}$ has $\mathbb{E}Z_i = p_i$ and all $\{Z_i\}_{i\in[n]}$ are independent. If $\mu := \mathbb{E}Z \geq \frac{2}{\epsilon^2}\log(\frac{2}{\delta})$ where $\delta, \epsilon \in (0, \frac{1}{2})$, then*

$$\Pr[Z \notin [(1-\epsilon)\mu, (1+\epsilon)\mu]] \leq \delta.$$

In other words, Corollary 2 says that if we are counting the number of random events $Z$ which either occur or do not (out of $n$ total events), but $\gg \log(\frac{1}{\delta})$ random events occur in expectation, then $Z$ concentrates very tightly around its expectation, independent of the number $n$. For example, Corollary 2 readily implies the following geometric aggregation technique.

**Lemma 4.** *Let $\delta \in (0,1)$, $R \geq 0$, and let $\mathcal{A}$ be a randomized algorithm returning $\mathbf{x} \in \mathbb{R}^d$ with $\mathbb{E}\left\|\mathbf{x} - \mathbf{x}^\star\right\|_2 \leq R$ for unknown $\mathbf{x}^\star \in \mathbb{R}^d$. There is an algorithm $\mathcal{A}'$ which calls $k$ independent copies of $\mathcal{A}$, and produces $\bar{\mathbf{x}} \in \mathbb{R}^d$ such that $\left\|\bar{\mathbf{x}} - \mathbf{x}^\star\right\|_2 \leq 9R$ with probability $\geq 1-\delta$, for $k = O(\log\frac{1}{\delta})$.*

*Proof.* Let $\mathcal{E}$ be the event that a point $\mathbf{x}$ satisfies $\left\|\mathbf{x} - \mathbf{x}^\star\right\|_2 \leq 3R$. By Markov's inequality (Fact 1), the output $\mathbf{x}_i$ of each independent run $i \in [k]$ satisfies $\mathcal{E}$ with probability $\geq \frac{2}{3}$. Therefore, Corollary 2 shows that for sufficiently large $k = O(\log\frac{1}{\delta})$, at least $\frac{3}{5}$ of the elements in $S := \{\mathbf{x}_i\}_{i\in[k]}$ will satisfy $\mathcal{E}$ with probability $\geq 1-\delta$. Condition on this happening for the rest of the proof.

We return any $\bar{\mathbf{x}} \in S$ such that $\left\|\bar{\mathbf{x}} - \mathbf{x}_i\right\|_2 \leq 6R$ for at least $\frac{3}{5}$ of the elements in $S$;[10] any point satisfying $\mathcal{E}$ is a valid $\bar{\mathbf{x}}$ by the triangle inequality. We claim $\left\|\bar{\mathbf{x}} - \mathbf{x}^\star\right\|_2 \leq 9R$. Suppose otherwise; then the triangle inequality shows $\left\|\mathbf{x}_i - \mathbf{x}^\star\right\|_2 \leq 3R$ and $\left\|\mathbf{x}_i - \bar{\mathbf{x}}\right\|_2 \leq 6R$ cannot both happen, but this contradicts the definition of $\bar{\mathbf{x}}$, since at most $\frac{2}{5}$ of the $\mathbf{x}_i \in S$ can have $\left\|\mathbf{x}_i - \bar{\mathbf{x}}\right\|_2 \leq 6R$. □

Lemma 4 shows we can boost expected distance guarantees to hold with high probability at an $O(\log\frac{1}{\delta})$ overhead in the number of runs, and a constant factor overhead in the distance bound. This is often useful for improving the guarantees of heavy-tailed stochastic optimization.

We defer a proof of Theorem 3 to Section 3, where we prove a generalization of it as Theorem 12. A proof of Theorem 3 can also be found in Section 2.3, [Ver18]; we mention that it follows by modifying the MGF method of Theorem 1 to use the inequality $(1-p) + p\exp(\lambda) \leq \exp(p(\exp(\lambda) - 1))$ for $p \in [0,1]$. In light of Remark 1, it is helpful to note that for Poisson random variables $Z$ with rate $\mu$, the corresponding Cramér transform has a closed-form

$$\psi^*(t) = \mu\left(\left(1 + \frac{t}{\mu}\right)\log\left(1 + \frac{t}{\mu}\right) - \frac{t}{\mu}\right).$$

Letting $t = \epsilon\mu$, we see that for small $\epsilon$, $\psi^*(t) \approx \epsilon^2$ (which can be thought of as the relatively sub-Gaussian regime), but for large $\epsilon$, $\psi^*(t) \approx \epsilon\log\epsilon$ (which falls more in line with the prediction

[10] Note that this can be performed without knowledge of $\mathbf{x}^\star$.

(5)). The actual proof of Theorem 3 can be reinterpreted by directly taking the Cramér transform of a sum of Bernoulli random variables, after scaling the problem by $\frac{1}{R}$.

**Remark 2.** *This section is best treated as a crash course or short review of the most useful scalar concentration inequalities. There are many other concentration inequalities which are good to know about (see [BLM13] for a comprehensive survey), some of which we briefly describe here.*

1. *Beyond sums, we can consider functions $Z = f(\{Z_i\}_{i\in[n]})$ which are additively stable in their inputs, i.e., for any fixed realization of the $\{Z_j\}_{j\neq i}$, the realization of $Z_i$ can only affect $Z$ by a bounded amount. This is known as the* bounded difference *property, and is captured by* McDiarmid's inequality *(a second-order generalization is the* Efron-Stein inequality*).*
2. *Each of Theorems 1, 2, and 3 also have generalizations to the* martingale *setting, where the random variable $Z$ is a sum of non-independent random variables $\{Z_i\}_{i\in[n]}$, but the conditional distributions $Z_i \mid \{Z_j\}_{j\in[i-1]}$ are appropriately behaved (e.g., sub-Gaussian or sub-gamma). These martingale variations are often known as* Azuma's inequalities*.*
3. *Instead of using the MGF to control tail behavior, we can sometimes obtain improvements by considering different potential functions, such as $p^{th}$ moments for large $p$. Such strategies result in bounds such as the* Khintchine inequality *or the* Marcinkiewicz–Zygmund inequality*.*
4. *Finally, one of the most commonly studied functions of a collection of random variables is their maximum (or supremum), which can be particularly challenging to control when the collection is infinite. Careful instantiations of net arguments (e.g., constructing chains of coverings, see Definition 2, Part II) are particularly useful in this setting. We defer an introduction to* maximal inequalities *to the excellent resource [vH16].*

# 2 Matrix analysis

We give an introduction to matrix analysis, which combined with the proof techniques in Section 1, will imply concentration bounds on random matrices. We emphasize that there are important techniques which matrix analysis inherits from scalar analysis, but one must be careful in general; as we will see, several basic inequalities which hold true for scalars are false for matrices.

## 2.1 Matrices as a vector space

The real $m \times n$ matrices, denoted $\mathbb{R}^{m\times n}$, are naturally identified with a vector space of dimension $mn$. We define the resulting inner product (the *Frobenius inner product*) of $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$ by

$$\langle \mathbf{A}, \mathbf{B}\rangle = \sum_{i\in[m], j\in[n]} \mathbf{A}_{ij}\mathbf{B}_{ij} = \mathrm{Tr}(\mathbf{A}^\top \mathbf{B}). \tag{6}$$

Correspondingly, $\mathbb{R}^{m\times n}$ is a normed space, where the relevant norm is the *Frobenius norm*:

$$\|\mathbf{A}\|_{\mathrm{F}} := \sqrt{\langle \mathbf{A}, \mathbf{A}\rangle} = \sqrt{\mathrm{Tr}\left(\mathbf{A}^\top \mathbf{A}\right)}.$$

Observe that $\|\mathbf{A}\|_{\mathrm{F}}$ is the entrywise $\ell_2$ norm, i.e. the $\ell_2$ norm of $\mathbf{A}$ as a vector in $\mathbb{R}^{mn}$. A particularly significant subspace of $\mathbb{R}^{d\times d}$ is the symmetric matrices, denoted $\mathbb{S}^{d\times d}$, which is a linear subspace of dimension $d + \binom{d}{2}$ (enforcing symmetry constraints). One reason for the elevated importance of $\mathbb{S}^{d\times d}$ is that it admits a convenient algebraic characterization via the spectral theorem.

**Theorem 4** (Spectral theorem)**.** *Let $\mathbf{M} \in \mathbb{R}^{d\times d}$. Then $\mathbf{M} \in \mathbb{S}^{d\times d}$ iff there is unitary*[11] *$\mathbf{U} \in \mathbb{R}^{d\times d}$ and diagonal $\mathbf{\Lambda} = \mathbf{diag}\,(\boldsymbol{\lambda}) \in \mathbb{R}^{d\times d}$ such that $\mathbf{M} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$. If $\mathbf{M} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ for unitary $\mathbf{U}$ and $\mathbf{\Lambda} = \mathbf{diag}\,(\boldsymbol{\lambda})$, then $\boldsymbol{\lambda}$ are the eigenvalues of $\mathbf{M}$ and the columns of $\mathbf{U}$ are the eigenvectors of $\mathbf{M}$.*

If $\mathbf{M}, \mathbf{N} \in \mathbb{S}^{d\times d}$ share a set of eigenvectors given by columns of the unitary matrix $\mathbf{U}$, then it is straightforward to check that $\mathbf{MN} = \mathbf{NM}$, i.e. $\mathbf{M}$ and $\mathbf{N}$ commute. Furthermore, we observe that Theorem 4 implies the trace of a symmetric matrix is the sum of its eigenvalues:

$$\mathrm{Tr}\,(\mathbf{M}) = \mathrm{Tr}\left(\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top\right) = \mathrm{Tr}\left(\mathbf{\Lambda}\mathbf{U}^\top\mathbf{U}\right) = \mathrm{Tr}(\mathbf{\Lambda}) \text{ for } \mathbf{M} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top \in \mathbb{S}^{d\times d}. \tag{7}$$

[11] Recall we say $\mathbf{U} \in \mathbb{R}^{d\times d}$ is *unitary* if $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_d$, i.e., $\mathbf{U}$ has orthonormal rows and columns.

Here we used the *cyclic property* of Tr, i.e., $\mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA})$ for all $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{B} \in \mathbb{R}^{n\times m}$ of compatible dimensions. Next, we denote the positive semidefinite $d \times d$ matrices by $\mathbb{S}^{d\times d}_{\succeq \mathbf{0}} \subset \mathbb{S}^{d\times d}$:

$$\mathbf{M} \in \mathbb{S}^{d\times d}_{\succeq \mathbf{0}} \iff \mathbf{v}^\top \mathbf{M}\mathbf{v} \geq 0 \text{ for all } \mathbf{v} \in \mathbb{R}^d. \tag{8}$$

By Theorem 4, $\mathbf{M} \in \mathbb{S}^{d\times d}$ satisfies $\mathbf{M} \in \mathbb{S}^{d\times d}_{\succeq \mathbf{0}}$ iff it has only nonnegative eigenvalues. Moreover, $\mathbf{M} \in \mathbb{S}^{d\times d}_{\succeq \mathbf{0}}$ iff there exists $\mathbf{A} \in \mathbb{R}^{n\times d}$ for some $n \in \mathbb{N}$ with $\mathbf{M} = \mathbf{A}^\top \mathbf{A}$. Similarly, the positive definite matrices are $\mathbb{S}^{d\times d}_{\succ \mathbf{0}}$ (i.e., all $\mathbf{M} \in \mathbb{S}^{d\times d}$ with only positive eigenvalues). We next mention a few additional interesting properties of $\mathbb{S}^{d\times d}_{\succeq \mathbf{0}}$. The first is a characterization via *Schur complements*.

**Fact 4.** *For a square matrix* $\mathbf{M}$ *partitioned into* $2 \times 2$ *blocks as follows:*

$$\mathbf{M} = \begin{pmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{pmatrix},$$

*if* $\mathbf{M}_{11}$ *and* $\mathbf{M}_{22}$ *are invertible, we define the* Schur complement *onto the top-left block by* $\mathrm{SC}_1(\mathbf{M}) := \mathbf{M}_{11} - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{21}$*, and the Schur complement onto the bottom-right block by* $\mathrm{SC}_2(\mathbf{M}) := \mathbf{M}_{22} - \mathbf{M}_{21}\mathbf{M}_{11}^{-1}\mathbf{M}_{12}$*. Then* $\mathbf{M}$ *is positive definite iff* $\mathrm{SC}_1(\mathbf{M}), \mathrm{SC}_2(\mathbf{M})$ *are positive definite.*[12]

We mention that one proof of Fact 4 uses that one can derive the Schur complement by considering a quadratic function $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{M}\mathbf{x}$ with respect to a partitioned variable $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$, and taking the partially-minimized function $g(\mathbf{x}_1) := \min_{\mathbf{x}_2} f(\mathbf{x}_1, \mathbf{x}_2)$, which turns out to be a quadratic in $\mathrm{SC}_1(\mathbf{M})$ if $\mathbf{M}$ is positive semidefinite; we defer more details to Appendix A.5.5, [BV04].

The nonnegativity structure of the set $\mathbb{S}^{d\times d}_{\succeq \mathbf{0}}$ also often finds use through the next result.

**Lemma 5.** *If* $\mathbf{M}, \mathbf{N} \in \mathbb{S}^{d\times d}_{\succeq \mathbf{0}}$*, we have* $\langle \mathbf{M}, \mathbf{N} \rangle \geq 0$.

*Proof.* Let $\mathbf{M} = \sum_{i\in[d]} \boldsymbol{\lambda}_i \mathbf{u}_i \mathbf{u}_i^\top$ and $\mathbf{N} = \sum_{i\in[d]} \boldsymbol{\lambda}'_i \mathbf{v}_i \mathbf{v}_i^\top$ where the $\{\mathbf{u}_i\}_{i\in[d]}$, $\{\mathbf{v}_i\}_{i\in[d]}$ are orthonormal sets of vectors, and $\{\boldsymbol{\lambda}_i\}_{i\in[d]}$, $\{\boldsymbol{\lambda}'_i\}_{i\in[d]}$ are all nonnegative. Then,

$$\begin{aligned} \langle \mathbf{M}, \mathbf{N} \rangle &= \left\langle \sum_{i\in[d]} \boldsymbol{\lambda}_i \mathbf{u}_i \mathbf{u}_i^\top, \sum_{i\in[d]} \boldsymbol{\lambda}'_i \mathbf{v}_i \mathbf{v}_i^\top \right\rangle = \sum_{i\in[d]} \sum_{j\in[d]} \boldsymbol{\lambda}_i \boldsymbol{\lambda}'_j \left\langle \mathbf{u}_i \mathbf{u}_i^\top, \mathbf{v}_j \mathbf{v}_j^\top \right\rangle \\ &= \sum_{i\in[d]} \sum_{j\in[d]} \boldsymbol{\lambda}_i \boldsymbol{\lambda}'_j \mathrm{Tr}\left(\mathbf{u}_i \mathbf{u}_i^\top \mathbf{v}_j \mathbf{v}_j^\top\right) = \sum_{i\in[d]} \sum_{j\in[d]} \boldsymbol{\lambda}_i \boldsymbol{\lambda}'_j \left\langle \mathbf{u}_i, \mathbf{v}_j \right\rangle^2 \geq 0. \end{aligned} \tag{9}$$

The second equality used that (6) is linear in its inputs, and the fourth used the cyclic property. □

There is a natural partial ordering $\preceq$ (called the *Löwner order*) induced on $\mathbb{S}^{d\times d}$, where for $\mathbf{M}, \mathbf{N} \in \mathbb{S}^{d\times d}$ we write $\mathbf{M} \preceq \mathbf{N}$ iff $\mathbf{N} - \mathbf{M} \in \mathbb{S}^{d\times d}_{\succeq \mathbf{0}}$ (respectively, $\mathbf{M} \prec \mathbf{N}$ iff $\mathbf{N} - \mathbf{M} \in \mathbb{S}^{d\times d}_{\succ \mathbf{0}}$). Note that not all matrices are comparable via $\preceq$. Moreover, Lemma 5 implies that if $\mathbf{N} \in \mathbb{S}^{d\times d}_{\succeq \mathbf{0}}$ and $\mathbf{M} \preceq \mathbf{M}'$,

$$\langle \mathbf{M}, \mathbf{N} \rangle \leq \langle \mathbf{M}', \mathbf{N} \rangle .$$

One other useful consequence of Theorem 4 is the existence of the *singular value decomposition*.

**Corollary 3** (Singular value decomposition)**.** *Let* $\mathbf{A} \in \mathbb{R}^{m\times n}$*. There exist unitary* $\mathbf{U} \in \mathbb{R}^{m\times m}$*,* $\mathbf{V} \in \mathbb{R}^{n\times n}$ *and* $\mathbf{\Sigma} \in \mathbb{R}^{m\times n}_{\geq 0}$ *where* $\mathbf{\Sigma}_{ij} = 0$ *for* $i \neq j$*, such that* $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$.

*Proof.* Let $\mathbf{A}$ have rank $r \leq \min(m, n)$. Let $\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ be the eigendecomposition of $\mathbf{A}\mathbf{A}^\top \in \mathbb{S}^{m\times m}_{\succeq \mathbf{0}}$ given by Theorem 4, with $\mathbf{\Lambda} = \mathbf{diag}(\boldsymbol{\lambda})$, and let $\{\mathbf{u}_i\}_{i\in[r]} \subset \mathbb{R}^m$ be the columns of $\mathbf{U}$ corresponding to nonzero entries in $\boldsymbol{\lambda}$. By definition, $\mathbf{A}\mathbf{A}^\top \mathbf{u}_i = \boldsymbol{\sigma}_i^2 \mathbf{u}_i$ for all $i \in [r]$, where $\boldsymbol{\sigma} := \sqrt{\boldsymbol{\lambda}}$ entrywise. Moreover, $\mathbf{A}^\top \mathbf{A}(\mathbf{A}^\top \mathbf{u}_i) = \boldsymbol{\sigma}_i^2 \mathbf{A}^\top \mathbf{u}_i$, and $\left\|\mathbf{A}^\top \mathbf{u}_i\right\|_2^2 = \boldsymbol{\sigma}_i^2$, so letting $\mathbf{v}_i := \frac{1}{\boldsymbol{\sigma}_i}\mathbf{A}^\top \mathbf{u}_i$ we see $\mathbf{v}_i$ is an unit-length eigenvector of $\mathbf{A}^\top \mathbf{A}$ for all $i \in [r]$. We also have by orthonormality of $\mathbf{U}$,

$$\mathbf{v}_i^\top \mathbf{v}_j = \frac{1}{\boldsymbol{\sigma}_i \boldsymbol{\sigma}_j} \mathbf{u}_i^\top \mathbf{A}\mathbf{A}^\top \mathbf{u}_j = \frac{\boldsymbol{\sigma}_i}{\boldsymbol{\sigma}_j} \mathbf{u}_i^\top \mathbf{u}_j = 0 \text{ for } i \neq j,$$

[12] This can be appropriately generalized to positive semidefinite matrices, where one takes care to handle kernels.

so the $\{\mathbf{v}_i\}_{i\in[r]}$ are orthonormal and can be completed to a unitary matrix $\mathbf{V}\in\mathbb{R}^{n\times n}$. Finally, letting $\mathbf{\Sigma}\in\mathbb{R}^{m\times n}$ have its first $r$ diagonal entries correspond to nonzero entries of $\boldsymbol{\sigma}$,[13]

$$\mathbf{V}=\mathbf{A}^\top\mathbf{U}\mathbf{\Sigma}^{-1}\implies\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top=\mathbf{A}^\top\implies\mathbf{A}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top.$$

□

We call the main diagonal elements of $\mathbf{\Sigma}$ given by Corollary 3, i.e., the set $\{\mathbf{\Sigma}_{ii}\}_{i\in[\min(m,n)]}$, the *singular values* of $\mathbf{A}$, and when $\mathbf{A}\in\mathbb{S}^{d\times d}_{\succeq\mathbf{0}}$, we note that they are the same as the eigenvalues of $\mathbf{A}$.

For a function $f:I\to\mathbb{R}$ where $I\subseteq\mathbb{R}$, and $\mathbf{M}\in\mathbb{S}^{d\times d}$ with eigenvalues all lying in $I$, we define a matrix function $f(\mathbf{M}):=\mathbf{U}f(\mathbf{\Lambda})\mathbf{U}^\top$ where $\mathbf{M}=\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ is the decomposition given by Theorem 4, and $f(\mathbf{\Lambda})$ is entrywise on the diagonal. Note that $f(\mathbf{M})$ always commutes with $\mathbf{M}$. We can check that when $f$ is a polynomial, this definition agrees with our intuition, since repeatedly applying $\mathbf{U}^\top\mathbf{U}=\mathbf{I}_d$ shows $\mathbf{M}^2=\mathbf{U}\mathbf{\Lambda}^2\mathbf{U}^\top$, and so on. We finally mention a few common matrix norms.

1. For $p\geq1$, we define the Schatten-$p$ norm of a matrix $\mathbf{A}\in\mathbb{R}^{m\times n}$, denoted $\|\mathbf{A}\|_p$, to be the $\ell_p$ norm of the *singular values* of $\mathbf{A}$. Our proof of Corollary 3 shows

$$\|\mathbf{A}\|_p^p=\mathrm{Tr}\left(\left(\mathbf{A}^\top\mathbf{A}\right)^{\frac{p}{2}}\right).$$

   We specifically reserve the names $\|\mathbf{A}\|_{\mathrm{tr}}:=\|\mathbf{A}\|_1$, $\|\mathbf{A}\|_{\mathrm{F}}:=\|\mathbf{A}\|_2$, and $\|\mathbf{A}\|_{\mathrm{op}}:=\|\mathbf{A}\|_\infty$.

2. As seen in Part IV, for $p,q\geq1$ we let the $p\to q$ operator norm of a matrix $\mathbf{A}\in\mathbb{R}^{m\times n}$ be defined by $\|\mathbf{A}\|_{p\to q}:=\max_{\|\mathbf{x}\|_p\leq1}\|\mathbf{A}\mathbf{x}\|_q$. One important fact is that $\|\mathbf{A}\|_{\mathrm{op}}=\|\mathbf{A}\|_{2\to2}$.

We denote the $i^{\text{th}}$ largest eigenvalue of $\mathbf{M}\in\mathbb{S}^{d\times d}$, and the $i^{\text{th}}$ largest singular value of $\mathbf{A}\in\mathbb{R}^{m\times n}$, by $\boldsymbol{\lambda}_i(\mathbf{M})$ and $\boldsymbol{\sigma}_i(\mathbf{A})$ respectively; note $\|\mathbf{A}\|_{\mathrm{op}}=\boldsymbol{\sigma}_1(\mathbf{A})=\sqrt{\boldsymbol{\lambda}_1(\mathbf{A}^\top\mathbf{A})}$ by the proof of Corollary 3.

## 2.2 Matrix inequalities

We now move to the topic of matrix inequalities, where we wish to establish relationships between matrices. Perhaps the simplest matrix inequality to establish is one which compares two commuting matrices, as then (after diagonalizing via Theorem 4) it reduces to multiple scalar inequalities.

**Lemma 6.** *Let $\mathbf{M}\in\mathbb{S}^{d\times d}$ have all its eigenvalues lying in $I\subset R$, and suppose $f:I\to\mathbb{R}$ and $g:I\to\mathbb{R}$ have $f(x)\leq g(x)$ for all $x\in I$. Then, $f(\mathbf{M})\preceq g(\mathbf{M})$.*

*Proof.* Write $\mathbf{M}=\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ by the eigendecomposition in Theorem 4, and note that by assumption,

$$f(\mathbf{\Lambda})-g(\mathbf{\Lambda})\preceq\mathbf{0}_d\implies f(\mathbf{\Lambda})\preceq g(\mathbf{\Lambda})$$

The claim follows upon multiplying by $\mathbf{U}$ on the left and $\mathbf{U}^\top$ on the right, which preserves (8). □

Next, when comparing eigenvalues between matrices, the following characterization is often helpful.

**Proposition 2** (Min-max eigenvalue theorem)**.** *Let $\mathbf{M}\in\mathbb{S}^{d\times d}$. For all $k\in[d]$,*

$$\lambda_k(\mathbf{M})=\max_{\substack{\mathbf{U}\in\mathbb{R}^{d\times k}\\ \mathbf{U}^\top\mathbf{U}=\mathbf{I}_k}}\min_{\substack{\mathbf{u}\in\mathrm{Span}(\mathbf{U})\\ \|\mathbf{u}\|_2=1}}\mathbf{u}^\top\mathbf{M}\mathbf{u}=\min_{\substack{\mathbf{U}\in\mathbb{R}^{d\times(d-k+1)}\\ \mathbf{U}^\top\mathbf{U}=\mathbf{I}_{d-k+1}}}\max_{\substack{\mathbf{u}\in\mathrm{Span}(\mathbf{U})\\ \|\mathbf{u}\|_2=1}}\mathbf{u}^\top\mathbf{M}\mathbf{u}.\tag{10}$$

Proposition 2, which is also sometimes called the *variational characterization* of eigenvalues, straightforwardly implies several famous matrix inequalities, such as the following.

**Corollary 4** (Cauchy interlacing theorem)**.** *For any $\mathbf{M}\in\mathbb{S}^{d\times d}$ and $\mathbf{V}\in\mathbb{R}^{d\times r}$ with $\mathbf{V}^\top\mathbf{V}=\mathbf{I}_r$,*

$$\boldsymbol{\lambda}_{d-r+k}\left(\mathbf{M}\right)\leq\boldsymbol{\lambda}_k\left(\mathbf{V}^\top\mathbf{M}\mathbf{V}\right)\leq\boldsymbol{\lambda}_k\left(\mathbf{M}\right)\ \textit{for all } k\in[r].$$

[13] Zero entries of $\mathbf{\Sigma}$ do not affect the proof, since they do not show up when expanding $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$.

*Proof.* Observe that when $\mathbf{U} \in \mathbb{R}^{n\times k}$ for some $n \geq k$, and $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_k$, any unit vector in $\mathrm{Span}(\mathbf{U})$ can be written as $\mathbf{U}\mathbf{v}$ for some $\mathbf{v} \in \mathbb{R}^k$ with $\|\mathbf{v}\|_2 = 1$. To see the upper bound, let $\mathbf{U} \in \mathbb{R}^{r\times k}$ with $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_k$ realize (10) for the definition of $\boldsymbol{\lambda}_k(\mathbf{V}^\top\mathbf{M}\mathbf{V})$. Proposition 2 then shows

$$\boldsymbol{\lambda}_k(\mathbf{V}^\top\mathbf{M}\mathbf{V}) = \min_{\substack{\mathbf{v}\in\mathbb{R}^k \\ \|\mathbf{v}\|_2=1}} \mathbf{v}^\top\mathbf{U}^\top\mathbf{V}^\top\mathbf{M}\mathbf{V}\mathbf{U}\mathbf{v}.$$

Note that $\|\mathbf{V}\mathbf{U}\mathbf{v}\|_2 = 1$ which can be seen by squaring both sides and using orthonormality of $\mathbf{U}$ and $\mathbf{V}$. Moreover, $\mathbf{W} := \mathbf{V}\mathbf{U}$ has $\mathbf{W} \in \mathbb{R}^{d\times k}$ and $\mathbf{W}^\top\mathbf{W} = \mathbf{I}_k$. We thus have

$$\min_{\substack{\mathbf{v}\in\mathbb{R}^k \\ \|\mathbf{v}\|_2=1}} \mathbf{v}^\top\mathbf{U}^\top\mathbf{V}^\top\mathbf{M}\mathbf{V}\mathbf{U}\mathbf{v} = \min_{\substack{\mathbf{v}\in\mathbb{R}^k \\ \|\mathbf{v}\|_2=1}} \mathbf{v}^\top\mathbf{W}^\top\mathbf{M}\mathbf{W}\mathbf{v} \leq \max_{\substack{\mathbf{W}\in\mathbb{R}^{d\times k} \\ \mathbf{W}^\top\mathbf{W}=\mathbf{I}_k}} \min_{\substack{\mathbf{v}\in\mathbb{R}^k \\ \|\mathbf{v}\|_2=1}} \mathbf{v}^\top\mathbf{W}^\top\mathbf{M}\mathbf{W}\mathbf{v} = \boldsymbol{\lambda}_k(\mathbf{M}).$$

The lower bound is similar: for some $\mathbf{U} \in \mathbb{R}^{d\times(r-k+1)}$ with orthonormal columns, and $\mathbf{W} := \mathbf{V}\mathbf{U}$,

$$\boldsymbol{\lambda}_k(\mathbf{V}^\top\mathbf{M}\mathbf{V}) = \max_{\substack{\mathbf{v}\in\mathbb{R}^k \\ \|\mathbf{v}\|_2=1}} \mathbf{v}^\top\mathbf{U}^\top\mathbf{V}^\top\mathbf{M}\mathbf{V}\mathbf{U}\mathbf{v} \geq \min_{\substack{\mathbf{W}\in\mathbb{R}^{d\times(r-k+1)} \\ \mathbf{W}^\top\mathbf{W}=\mathbf{I}_{r-k+1}}} \max_{\substack{\mathbf{v}\in\mathbb{R}^k \\ \|\mathbf{v}\|_2=1}} \mathbf{v}^\top\mathbf{W}^\top\mathbf{M}\mathbf{W}\mathbf{v} = \boldsymbol{\lambda}_{d-r+k}(\mathbf{M}).$$

□

When $r = d - 1$, observe that Corollary 4 says the eigenvalues of $\mathbf{V}\mathbf{M}\mathbf{V}^\top$ *interlace* those of $\mathbf{M}$.

We remark that the matrix $\mathbf{P} = \mathbf{V}\mathbf{V}^\top$ is often called the *orthogonal projection matrix* or *orthogonal projector* onto the subspace spanned by $\mathbf{V}$. All eigenvalues of $\mathbf{P}$ are in $\{0, 1\}$, and further any $\mathbf{P} \in \mathbb{S}^{d\times d}$ with eigenvalues all in $\{0, 1\}$ is an orthogonal projector onto the subspace spanned by its nonzero eigenvalues, and satisfies $\mathbf{P}^2 = \mathbf{P}$.[14]

Proposition 2 also implies Weyl's inequality, a famous result in matrix perturbation analysis.

**Corollary 5** (Weyl's inequality)**.** *Let $\mathbf{M}, \mathbf{N} \in \mathbb{S}^{d\times d}$. For all $k \in [d]$,*

$$\boldsymbol{\lambda}_k(\mathbf{M}) + \boldsymbol{\lambda}_d(\mathbf{N}) \leq \boldsymbol{\lambda}_k(\mathbf{M} + \mathbf{N}) \leq \boldsymbol{\lambda}_k(\mathbf{M}) + \boldsymbol{\lambda}_1(\mathbf{N}).$$

*Proof.* The upper bound follows from Proposition 2, which shows

$$\begin{aligned}\boldsymbol{\lambda}_k(\mathbf{M} + \mathbf{N}) &= \max_{\substack{\mathbf{U}\in\mathbb{R}^{d\times k} \\ \mathbf{U}^\top\mathbf{U}=\mathbf{I}_k}} \min_{\substack{\mathbf{u}\in\mathrm{Span}(\mathbf{U}) \\ \|\mathbf{u}\|_2=1}} \mathbf{u}^\top(\mathbf{M} + \mathbf{N})\mathbf{u} \\ &\leq \max_{\substack{\mathbf{U}\in\mathbb{R}^{d\times k} \\ \mathbf{U}^\top\mathbf{U}=\mathbf{I}_k}} \min_{\substack{\mathbf{u}\in\mathrm{Span}(\mathbf{U}) \\ \|\mathbf{u}\|_2=1}} \mathbf{u}^\top\mathbf{M}\mathbf{u} + \max_{\|\mathbf{u}\|_2=1} \mathbf{u}^\top\mathbf{N}\mathbf{u} = \boldsymbol{\lambda}_k(\mathbf{M}) + \boldsymbol{\lambda}_1(\mathbf{N}).\end{aligned}$$

The lower bound follows similarly, where we use the other characterization in (10). □

Finally, we mention one additional consequence of Proposition 2.

**Corollary 6.** *Let $f : I \to \mathbb{R}$ be monotone nondecreasing for $I \subseteq \mathbb{R}$, and let $\mathbf{M}, \mathbf{N} \in \mathbb{S}^{d\times d}$ have all their eigenvalues in $I$. Then if $\mathbf{M} \preceq \mathbf{N}$, we have $\mathrm{Tr}(f(\mathbf{M})) \leq \mathrm{Tr}(f(\mathbf{N}))$.*

*Proof.* For $k \in [d]$, Proposition 2 and the proof of Corollary 4 show $\boldsymbol{\lambda}_k(\mathbf{M}) \leq \boldsymbol{\lambda}_k(\mathbf{N})$ for all $k \in [d]$:

$$\boldsymbol{\lambda}_k(\mathbf{M}) = \min_{\substack{\mathbf{v}\in\mathbb{R}^k \\ \|\mathbf{v}\|_2=1}} \mathbf{v}^\top\mathbf{U}^\top\mathbf{M}\mathbf{U}\mathbf{v} \leq \min_{\substack{\mathbf{v}\in\mathbb{R}^k \\ \|\mathbf{v}\|_2=1}} \mathbf{v}^\top\mathbf{U}^\top\mathbf{N}\mathbf{U}\mathbf{v} \leq \boldsymbol{\lambda}_k(\mathbf{N}), \text{ for some } \mathbf{U} \in \mathbb{R}^{d\times k} \text{ with } \mathbf{U}^\top\mathbf{U} = \mathbf{I}_k.$$

This implies the claim via $\sum_{k\in[d]} f(\boldsymbol{\lambda}_k(\mathbf{M})) \leq \sum_{k\in[d]} f(\boldsymbol{\lambda}_k(\mathbf{N}))$, where we recall (7). □

Upon seeing Corollary 6, one may hope that an even stronger fact is true: i.e., if $f$ is monotone nondecreasing, then $f(\mathbf{M}) \preceq f(\mathbf{N})$. Note that if this were true, Corollary 6 would follow by applying Lemma 5 with $\mathbf{M} \leftarrow f(\mathbf{N}) - f(\mathbf{M})$ and $\mathbf{N} \leftarrow \mathbf{I}_d$. Unfortunately, this stronger fact is false in many cases. For example, $\mathbf{A} \preceq \mathbf{B}$ does not imply $\mathbf{A}^2 \preceq \mathbf{B}^2$, as witnessed by the example

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \ \mathbf{B} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}, \ \text{such that } \det(\mathbf{B}^2 - \mathbf{A}^2) = \det \begin{pmatrix} 3 & 1 \\ 1 & 0 \end{pmatrix} < 0.$$

We define the following properties of functions acting on symmetric matrices.

[14] This has the nice interpretation that projecting onto a subspace twice is the same as projecting once.

**Definition 3.** *Let $f : \mathbb{S}_I^{d\times d} \to \mathbb{S}^{d\times d}$ be an operator function identified with a scalar function $f : I \to \mathbb{R}$, where $\mathbb{S}_I^{d\times d}$ is the subset of $\mathbb{S}^{d\times d}$ with eigenvalues contained in $I \subseteq \mathbb{R}$.*

1. *We say $f$ is operator monotone if $f(\mathbf{M}) \preceq f(\mathbf{N})$ for any $\mathbf{M}, \mathbf{N} \in \mathbb{S}_I^{d\times d}$ with $\mathbf{M} \preceq \mathbf{N}$.*
2. *We say $f$ is operator convex if $f((1-\lambda)\mathbf{M}+\lambda\mathbf{M}') \preceq (1-\lambda)f(\mathbf{M})+\lambda f(\mathbf{M}')$ for any $\mathbf{M}, \mathbf{M}' \in \mathbb{S}_I^{d\times d}$ and $\lambda \in (0,1)$. If $-f$ is operator convex, then we say $f$ is operator concave.*

The next claim summarizes some well-known cases where the operator function does inherit the corresponding scalar property; for a proof, we refer the reader to Lecture 5 of [Lee21].

**Theorem 5** (Löwner-Heinz)**.** *Let $f : \mathbb{R}_{>0} \to \mathbb{R}$ be identified with its operator counterpart.*

1. *If $f(x) = x^p$ for $p \in [-1, 0]$, $f$ is operator convex and $-f$ is operator monotone.*
2. *If $f(x) = x^p$ for $p \in [0, 1]$, or $f(x) = \log x$, $f$ is operator monotone and operator concave.*
3. *If $f(x) = x^p$ for $p \in [1, 2]$, or $f(x) = x \log x$, $f$ is operator convex.*

We mention that the non-polynomial $f$ in Theorem 5 inherit properties from polynomials via

$$\log x = \lim_{p\to 0} \frac{x^p - 1}{p}, \; x \log x = \lim_{p\to 1} \frac{x^p - x}{p-1}.$$

Finally, we mention two themes which often arise when proving matrix inequalities. The first theme is that inequalities are sometimes tight when eigenspaces of matrices are aligned, i.e., when matrices commute. This is based on the fact that for $\mathbf{M}, \mathbf{N} \in \mathbb{S}^{d\times d}$ with fixed eigenvalues $\{\boldsymbol{\lambda}_i\}_{i\in[d]}$ and $\{\boldsymbol{\lambda}'_i\}_{i\in[d]}$, $\langle \mathbf{M}, \mathbf{N}\rangle$ attains both its maximum and minimum over the eigenvectors of $\mathbf{N}$ when $\mathbf{M}$ and $\mathbf{N}$ commute. In other words, extremal inner products are achieved by commuting matrices.

**Theorem 6** (von Neumann trace inequality)**.** *Let $\mathbf{M}, \mathbf{N} \in \mathbb{S}^{d\times d}$ have eigendecompositions $\mathbf{M} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$ and $\mathbf{V}\boldsymbol{\Lambda}'\mathbf{V}^\top$ given by Theorem 4, where $\boldsymbol{\Lambda} = \mathbf{diag}(\boldsymbol{\lambda})$ and $\boldsymbol{\Lambda}' = \mathbf{diag}(\boldsymbol{\lambda}')$, and we assume that the $\{\boldsymbol{\lambda}_i\}_{i\in[d]}$ and $\{\boldsymbol{\lambda}'_i\}_{i\in[d]}$ are sorted in nondecreasing order. Then*

$$\sum_{i\in[d]} \boldsymbol{\lambda}_i \boldsymbol{\lambda}'_{d+1-i} \le \langle \mathbf{M}, \mathbf{N}\rangle \le \sum_{i\in[d]} \boldsymbol{\lambda}_i \boldsymbol{\lambda}'_i.$$

*The upper and lower bounds respectively hold if $\mathbf{U} = \mathbf{V}$ and if columns of $\mathbf{U}$, $\mathbf{V}$ are reversed.*

*Proof.* Observe that by the calculation (9), we have $\langle \mathbf{M}, \mathbf{N}\rangle = \langle \mathbf{S}, \mathbf{T}\rangle$, where

$$\mathbf{S}_{ij} = \langle \mathbf{u}_i, \mathbf{v}_j\rangle^2, \; \mathbf{T}_{ij} = \boldsymbol{\lambda}_i \boldsymbol{\lambda}'_j \text{ for all } (i,j) \in [d] \times [d].$$

Because $\mathbf{U}, \mathbf{V}$ are unitary, we have $\sum_{i\in[d]} \mathbf{S}_{ij} = \|\mathbf{U}^\top \mathbf{v}_j\|^2 = \|\mathbf{v}_j\|_2^2 = 1$ for all $j \in [d]$, and similarly $\sum_{j\in[d]} \mathbf{S}_{ij} = 1$ for all $i \in [d]$. This shows that $\mathbf{S}$ is a *doubly-stochastic matrix*, i.e., it is a nonnegative $d \times d$ matrix with all rows and columns summing to 1. So, we have

$$\begin{aligned} \langle \mathbf{M}, \mathbf{N}\rangle = \langle \mathbf{S}, \mathbf{T}\rangle \in \left[\langle \mathbf{S}^-, \mathbf{T}\rangle, \langle \mathbf{S}^+, \mathbf{T}\rangle\right], \\ \text{where } \mathbf{S}^- \in \operatorname{argmin}_{\mathbf{S}\in S} \langle \mathbf{S}, \mathbf{T}\rangle, \; \mathbf{S}^+ \in \operatorname{argmax}_{\mathbf{S}\in S} \langle \mathbf{S}, \mathbf{T}\rangle. \end{aligned} \tag{11}$$

It is well-known that the extremal points of $S$, the set of doubly-stochastic matrices, are the *permutation matrices*,[15] i.e., matrices in $\{0,1\}^{d\times d}$ with one entry per row and column equalling 1, and we can check that $S$ is convex. Moreover, $\langle \mathbf{S}, \mathbf{T}\rangle$ is a linear function in $\mathbf{S}$, so both it and its negation are convex in $\mathbf{S}$. Therefore, combining Lemma 3, Part I with (11) shows

$$\sum_{i\in[d]} \boldsymbol{\lambda}_i \boldsymbol{\lambda}'_{d+1-i} = \min_{\pi\in P} \sum_{i\in[d]} \boldsymbol{\lambda}_i \boldsymbol{\lambda}'_{\pi(i)} \le \langle \mathbf{M}, \mathbf{N}\rangle \le \max_{\pi\in P} \sum_{i\in[d]} \boldsymbol{\lambda}_i \boldsymbol{\lambda}'_{\pi(i)} = \sum_{i\in[d]} \boldsymbol{\lambda}_i \boldsymbol{\lambda}'_i.$$

where $P$ is the set of permutations $\pi : [d] \to [d]$, and we used the *rearrangement inequality* which says the minimizing and maximizing permutations are the ones which agree with or fully invert from the original ordering. This shows both claims by considering the extremal arguments. □

[15] This is the Birkhoff-von Neumann theorem, and follows from integrality of the bipartite matching polytope, since there is a bijection between doubly-stochastic matrices and perfect fractional bipartite matchings.

**Remark 3.** *Theorem 6 extends to pairs of asymmetric, complex square matrices* $(\mathbf{M}, \mathbf{N})$*, where the eigenvalues in the inequalities are replaced with singular values [vN37].*

One consequence of Theorem 6 (and its generalization in Remark 3) is a Cauchy-Schwarz inequality for all *unitarily-invariant norms* on square matrices. We say that a norm $\|\cdot\|$ on $\mathbb{R}^{m\times n}$ is unitarily-invariant if, for any unitary $\mathbf{U} \in \mathbb{R}^{m\times m}$, $\mathbf{V} \in \mathbb{R}^{n\times n}$, and any $\mathbf{M} \in \mathbb{R}^{d\times d}$, we have $\|\mathbf{M}\| = \left\|\mathbf{U}^\top \mathbf{M}\mathbf{V}\right\|$. This means $\|\cdot\|$ must be a function of only the singular values. For example, the Schatten-$p$ norms are unitarily-invariant. Because Theorem 6 says the maximizing arguments align eigenspaces, we can apply scalar Cauchy-Schwarz inequalities to matrices. Namely,

$$\langle \mathbf{M}, \mathbf{N}\rangle \le \|\mathbf{M}\|\,\|\mathbf{N}\|_* \text{ for all unitarily-invariant norms } \|\cdot\|,\ \mathbf{M}, \mathbf{N} \in \mathbb{R}^{d\times d}. \tag{12}$$

When $\|\cdot\|$ and its dual are Schatten norms, (12) recovers a matrix variant of Hölder's inequality.

The second main theme is looking for ways to disentangle products of matrices into portions which sort the same matrix with itself, e.g. for any $\mathbf{M}, \mathbf{N} \in \mathbb{S}^{d\times d}$,

$$\operatorname{Tr}(\mathbf{MNMN}) \le \operatorname{Tr}(\mathbf{M}^2\mathbf{N}^2). \tag{13}$$

The above inequality follows by applying (12) with the self-dual norm $\|\cdot\|_{\mathrm{F}}$, since $\|\mathbf{MN}\|_{\mathrm{F}}^2 = \operatorname{Tr}(\mathbf{NM}^2\mathbf{N}) = \operatorname{Tr}(\mathbf{M}^2\mathbf{N}^2)$ by the cyclic property. We demonstrate this second theme by giving a few exemplar building blocks of this disentangling strategy that often arise in proofs. The first generalizes a special case of (13), and recovers it when $\mathbf{N} \in \mathbb{S}^{d\times d}_{\succeq \mathbf{0}}$ by setting $\mathbf{M} \leftarrow \mathbf{M}^2$.

**Lemma 7.** *Let* $\mathbf{M}, \mathbf{N} \in \mathbb{S}^{d\times d}_{\succeq \mathbf{0}}$ *and* $\alpha \in (0, 1)$*. Then*

$$\operatorname{Tr}(\mathbf{M}^\alpha \mathbf{N}\mathbf{M}^{1-\alpha}\mathbf{N}) \le \operatorname{Tr}(\mathbf{MN}^2).$$

*Proof.* Let $f(\alpha) := \operatorname{Tr}(\mathbf{M}^\alpha \mathbf{N}\mathbf{M}^{1-\alpha}\mathbf{N})$ and note $f$ is symmetric about $\frac{1}{2}$ in the range $[0, 1]$. We claim that for all $\alpha \in [0, \frac{1}{2}]$, we have $f(\alpha) \le \frac{1}{2}f(0) + \frac{1}{2}f(2\alpha)$. This implies that $f$ is convex in $[0, \frac{1}{2}]$, so it is maximized at either $\alpha = 0$ or $\alpha = \frac{1}{2}$, and hence 0 is a maximizer since $f(\frac{1}{2}) \le \frac{1}{2}f(0) + \frac{1}{2}f(1) = f(0)$. To see our claim $f(\alpha) \le \frac{1}{2}f(0) + \frac{1}{2}f(2\alpha)$, consider the matrices

$$\mathbf{A} := \begin{pmatrix} \mathbf{N} & -\mathbf{N}^{\frac{1}{2}}\mathbf{M}^\alpha\mathbf{N}^{\frac{1}{2}} \\ -\mathbf{N}^{\frac{1}{2}}\mathbf{M}^\alpha\mathbf{N}^{\frac{1}{2}} & \mathbf{N}^{\frac{1}{2}}\mathbf{M}^{2\alpha}\mathbf{N}^{\frac{1}{2}} \end{pmatrix} = \begin{pmatrix} \mathbf{N}^{\frac{1}{2}} & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{N}^{\frac{1}{2}} \end{pmatrix} \begin{pmatrix} \mathbf{I}_d & -\mathbf{M}^\alpha \\ -\mathbf{M}^\alpha & \mathbf{M}^{2\alpha} \end{pmatrix} \begin{pmatrix} \mathbf{N}^{\frac{1}{2}} & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{N}^{\frac{1}{2}} \end{pmatrix},$$

$$\mathbf{B} := \begin{pmatrix} \mathbf{N}^{\frac{1}{2}}\mathbf{M}\mathbf{N}^{\frac{1}{2}} & -\mathbf{N}^{\frac{1}{2}}\mathbf{M}^{1-\alpha}\mathbf{N}^{\frac{1}{2}} \\ -\mathbf{N}^{\frac{1}{2}}\mathbf{M}^{1-\alpha}\mathbf{N}^{\frac{1}{2}} & \mathbf{N}^{\frac{1}{2}}\mathbf{M}^{1-2\alpha}\mathbf{N}^{\frac{1}{2}} \end{pmatrix} = \begin{pmatrix} \mathbf{N}^{\frac{1}{2}} & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{N}^{\frac{1}{2}} \end{pmatrix} \begin{pmatrix} \mathbf{M} & -\mathbf{M}^{1-\alpha} \\ -\mathbf{M}^{1-\alpha} & \mathbf{M}^{1-2\alpha} \end{pmatrix} \begin{pmatrix} \mathbf{N}^{\frac{1}{2}} & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{N}^{\frac{1}{2}} \end{pmatrix}.$$

Note that $\mathbf{A}$ and $\mathbf{B}$ are both in $\mathbb{S}^{2d\times 2d}_{\succeq \mathbf{0}}$: the middle matrices in each product are positive semidefinite by Fact 4, and the outer matrices preserve the definition (8). Therefore, rearranging the inequality $\langle \mathbf{A}, \mathbf{B}\rangle \ge 0$ (which follows from Lemma 5) after expanding by blocks proves the claim. □

Lemma 7 was first shown (to our knowledge) by [Nes07], and has been rediscovered in a number of different settings [Eld13, ZLO16]; in the latter work, it was termed an *extended Lieb-Thirring inequality* due to its relationship with a classical inequality of [LT76]. We summarize (without proof) the *Lieb-Thirring inequality* of [LT76] below, which also generalizes (13) when $\mathbf{M}, \mathbf{N} \in \mathbb{S}^{d\times d}_{\succeq \mathbf{0}}$.

**Proposition 3.** *Let* $\mathbf{M}, \mathbf{N} \in \mathbb{S}^{d\times d}_{\succeq \mathbf{0}}$ *and* $p \ge 1$*. Then* $\operatorname{Tr}((\mathbf{MN})^p) \le \operatorname{Tr}(\mathbf{M}^p\mathbf{N}^p)$.

Both Lemma 7 and Proposition 3 can be viewed as formalizations of our earlier intuition that "disentangling increases matrix products." Note that in both cases, the larger quantity in each inequality groups together copies of the same matrix. By using limiting arguments, these disentangling tools yield powerful conclusions such as the Golden-Thompson inequality [Gol65, Tho65].

**Theorem 7** (Golden-Thompson inequality)**.** *Let* $\mathbf{M}, \mathbf{N} \in \mathbb{S}^{d\times d}$*. Then*

$$\operatorname{Tr}(\exp(\mathbf{M} + \mathbf{N})) \le \operatorname{Tr}(\exp(\mathbf{M})\exp(\mathbf{N})). \tag{14}$$

*Proof.* We only provide a brief proof sketch here, deferring more details to [Tao10]. By using Taylor series expansions, we first note that (14) is reminiscent of our earlier disentangling conclusions. For example, the third-order Taylor expansions of both sides of (14) agree, and the fourth-order terms turn out to be comparable via (13). More generally, the left-hand side of (14) tends to

interleave copies of $\mathbf{M}, \mathbf{N}$ by expanding $(\mathbf{M}+\mathbf{N})^p$, whereas these copies are fully disentangled on the right-hand side. We relate our disentangling inequality Proposition 3 to (14) by observing

$$\lim_{p\to\infty}\left(\exp\left(\frac{1}{p}\mathbf{M}\right)\exp\left(\frac{1}{p}\mathbf{N}\right)\right)^p = \exp\left(\mathbf{M}+\mathbf{N}\right),$$

which follows since the first-order Taylor expansions of both sides agree, and all the higher-order terms vanish as $p\to\infty$. Now, let $\mathbf{A}\leftarrow\exp(\frac{1}{p}\mathbf{M})$ and $\mathbf{B}\leftarrow\exp(\frac{1}{p}\mathbf{N})$, so that the right-hand side of (14) is $\mathrm{Tr}(\mathbf{A}^p\mathbf{B}^p)$ and the left-hand side is the expression in the above limit. The conclusion follows by sending $p\to\infty$, and applying Proposition 3 with $\mathbf{A},\mathbf{B}$ in place of $\mathbf{M},\mathbf{N}$. □

## 2.3 Matrix calculus

In this section, we develop some basic rules for matrix calculus, specialized to convex functions of symmetric matrices. Our development is based on the works [Lew96, LS01], which heavily rely on the theory of convex conjugates from Part III, combined with Theorem 6. Specifically, consider an operator function $f:\mathbb{S}_I^{d\times d}\to\mathbb{R}$. We say $f$ is a *spectral function* if $f$ is only a function of the eigenvalues of its input, i.e. for all unitary $\mathbf{U}$ and $\mathbf{M}\in\mathbb{S}_I^{d\times d}$, $f(\mathbf{U}^\top\mathbf{M}\mathbf{U})=f(\mathbf{M})$. For example, the unitarily-invariant norms mentioned in (12) are spectral functions. When $f$ is a spectral function, we identify it with a vector counterpart $f_{\mathrm{vec}}:I^d\to\mathbb{R}$ mapping a set of eigenvalues to a scalar.

Now suppose $f_{\mathrm{vec}}$ is convex, closed, and proper, and consider the definition of the conjugate of $f$:

$$f^*(\mathbf{N}) := \max_{\mathbf{M}\in\mathbb{S}_I^{d\times d}} \langle\mathbf{M},\mathbf{N}\rangle - f(\mathbf{M}).$$

Because it is a spectral function, $f$ is invariant under unitary transformations. Moreover, Theorem 6 says for fixed eigenvalues the inner product is maximized when $\mathbf{M}$ commutes with $\mathbf{N}$. Rewriting the above expression in terms of diagonal matrices, we have shown $f^*(\mathbf{N})$ is simply $f_{\mathrm{vec}}^*$ applied to the eigenvalues of $\mathbf{N}$. A variation of this argument shows that when $f, f_{\mathrm{vec}}$ are differentiable, $\nabla f(\mathbf{M})$ commutes with $\mathbf{M}$ through the characterization of gradients as maximizing arguments in conjugates (Fact 1, Part III). The following summary of these observations is proven in [Lew96].

**Theorem 8** (Lewis's theorem)**.** *Let $f_{\mathrm{vec}}:\mathbb{R}^d\to\mathbb{R}$ be permutation-invariant,[16] convex, closed, and proper. Then $f:\mathbb{S}^{d\times d}\to\mathbb{R}$ identified with $f_{\mathrm{vec}}$ is permutation-invariant, convex, closed, and proper. Further, for $\mathbf{M},\mathbf{N}\in\mathbb{S}_{\succeq\mathbf{0}}^{d\times d}$ with eigendecompositions $\mathbf{M}=\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, $\mathbf{N}=\mathbf{V}\mathbf{\Lambda}'\mathbf{V}^\top$, where $\mathbf{\Lambda}=\mathbf{diag}\left(\boldsymbol{\lambda}\right)$, $\mathbf{\Lambda}'=\mathbf{diag}\left(\boldsymbol{\lambda}'\right)$, we have $\mathbf{N}\in\partial f(\mathbf{M})$ iff $\boldsymbol{\lambda}'\in\partial f_{\mathrm{vec}}(\boldsymbol{\lambda})$, $\mathbf{U}=\mathbf{V}$.*

In other words, if we know how to characterize the subdifferential structure of $f_{\mathrm{vec}}$, Theorem 8 tells us the subdifferential structure of $f$ as well. A common example of a spectral function satisfying the conditions of Theorem 8 is the trace of a map from symmetric matrices to symmetric matrices, e.g., those in Corollary 6.[17] Applying Theorem 8 then gives useful conclusions such as the following:

$$\nabla\left(\frac{1}{2}\left\|\mathbf{M}\right\|_{\mathrm{F}}^2\right)=\nabla\left(\mathrm{Tr}\left(\frac{1}{2}\mathbf{M}^2\right)\right)=\mathbf{M},\ \nabla\left(\mathrm{Tr}\exp\left(\mathbf{M}\right)\right)=\exp\left(\mathbf{M}\right).$$

Finally, we mention a formula for the Hessian of convex spectral functions derived in [LS01]. We give an example of how to perform computations involving Theorem 9 in a following lecture.

**Theorem 9** (Lewis-Sendov formula)**.** *Let $f_{\mathrm{vec}}:\mathbb{R}^d\to\mathbb{R}$ be permutation-invariant. Then the operator function $f:\mathbb{S}^{d\times d}\to\mathbb{R}$ identified with $f_{\mathrm{vec}}$ is permutation-invariant. Moreover, for any $\mathbf{M}\in\mathbb{S}_{\succeq\mathbf{0}}^{d\times d}$ with eigendecomposition $\mathbf{M}=\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, where $\mathbf{\Lambda}=\mathbf{diag}\left(\boldsymbol{\lambda}\right)$, $f$ is twice-differentiable at $\mathbf{M}$ iff $f_{\mathrm{vec}}$ is twice-differentiable at $\boldsymbol{\lambda}$. In this case, $\nabla^2 f(\mathbf{M})$ operates on $\mathbf{N}\in\mathbb{S}^{d\times d}$ via the following formula, where $\circ$ denotes entrywise multiplication:*

$$\nabla^2 f(\mathbf{M})[\mathbf{N}]=\mathbf{U}\left(\mathbf{diag}\left(\left\{\frac{\partial^2}{\partial\lambda_i^2}f_{\mathrm{vec}}(\boldsymbol{\lambda})\cdot\widetilde{\mathbf{N}}_{ii}\right\}_{i\in[d]}\right)+\mathbf{A}\circ\widetilde{\mathbf{N}}\right)\mathbf{U}^\top,$$

$$\text{where } \mathbf{A}_{ij}=\begin{cases}0 & i=j\\ \nabla_{ii}^2 f_{\mathrm{vec}}(\boldsymbol{\lambda})-\nabla_{ij}^2 f_{\mathrm{vec}}(\boldsymbol{\lambda}) & i\neq j,\ \boldsymbol{\lambda}_i=\boldsymbol{\lambda}_j\\ \frac{\nabla_i f_{\mathrm{vec}}(\boldsymbol{\lambda})-\nabla_j f_{\mathrm{vec}}(\boldsymbol{\lambda})}{\boldsymbol{\lambda}_i-\boldsymbol{\lambda}_j} & i\neq j,\ \boldsymbol{\lambda}_i\neq\boldsymbol{\lambda}_j\end{cases}\text{ for } (i,j)\in[d]\times[d],\ \widetilde{\mathbf{N}}:=\mathbf{U}^\top\mathbf{N}\mathbf{U}.$$

[16] This condition comes up when fully characterizing the equality cases of Theorem 6.

[17] Such functions are automatically permutation-invariant, due to permutation-invariance of the equality (7).

# 3 Matrix concentration

The study of matrix concentration inequalities develops tools for bounding statistics of random matrices. Perhaps the most common statistics of interest are the largest and smallest eigenvalues. For example, if we want to understand the deviation of a sum of random, mean-zero symmetric matrices from its expectation, a natural measure for deviation is the operator norm of the sum (i.e. the largest eigenvalue by magnitude). Towards this end, we show how Theorems 2 and 3 can be generalized to the matrix setting, resulting in the matrix Bernstein bound of [Tro11] and matrix Chernoff bound of [AW02]. These matrix concentration results are tremendously applicable to the analysis of randomized algorithms in numerical linear algebra, as we will see in future lectures. Our exposition closely follows strategies outlined in the excellent resource [Tro15].

## 3.1 Matrix MGF method

To begin, we give a suitable extension of the MGF method in Theorem 1, and as mentioned in Remark 1. Our goal is to control a sum of random matrices through the following basic claim.

**Lemma 8.** *Let $\mathbf{Z} \in \mathbb{S}^{d\times d}$ be a random matrix. Then,*[18]

$$\Pr\left[\boldsymbol{\lambda}_1\left(\mathbf{Z}\right) \geq t\right] \leq \inf_{\theta>0} \exp\left(-\theta t\right) \mathbb{E}\left[\operatorname{Tr}\exp\left(\theta\mathbf{Z}\right)\right],$$
$$\Pr\left[\boldsymbol{\lambda}_d\left(\mathbf{Z}\right) \leq t\right] \leq \inf_{\theta<0} \exp\left(-\theta t\right) \mathbb{E}\left[\operatorname{Tr}\exp\left(\theta\mathbf{Z}\right)\right].$$

*Proof.* We start with the upper tail bound, following the strategy in Theorem 1: for $\theta > 0$,

$$\begin{aligned}\Pr\left[\boldsymbol{\lambda}_1\left(\mathbf{Z}\right) \geq t\right] &= \Pr\left[\exp\left(\theta\boldsymbol{\lambda}_1\left(\mathbf{Z}\right)\right) \geq \exp\left(\theta t\right)\right] \\ &\leq \exp\left(-\theta t\right)\mathbb{E}\left[\exp\left(\theta\boldsymbol{\lambda}_1\left(\mathbf{Z}\right)\right)\right] \leq \exp\left(-\theta t\right)\mathbb{E}\left[\operatorname{Tr}\exp(\theta\mathbf{Z})\right].\end{aligned}$$

The first inequality was Markov's (Fact 1), and the second used that $\boldsymbol{\lambda}_1(\exp(\theta\mathbf{Z})) = \exp(\theta\boldsymbol{\lambda}_1(\mathbf{Z}))$ by nonnegativity of $\theta$ and the definition of the matrix exponential. The tail bound then holds because we had the freedom of choosing $\theta$. Similarly, for the lower tail, for $\theta < 0$,

$$\begin{aligned}\Pr\left[\boldsymbol{\lambda}_d\left(\mathbf{Z}\right) \leq t\right] &= \Pr\left[\exp\left(\theta\boldsymbol{\lambda}_d(\mathbf{Z})\right) \geq \exp\left(\theta t\right)\right] \\ &\leq \exp\left(-\theta t\right)\mathbb{E}\left[\exp\left(\theta\boldsymbol{\lambda}_d(\mathbf{Z})\right)\right] \leq \exp\left(-\theta t\right)\mathbb{E}\left[\operatorname{Tr}\exp\left(\theta\mathbf{Z}\right)\right].\end{aligned}$$

□

In the original MGF method, when $Z$ was a random scalar we controlled $\mathbb{E}[\exp(\theta Z)]$ for different values of $\theta$. Lemma 8 suggests that, analogously, a natural matrix potential to track is $\mathbb{E}\operatorname{Tr}\exp(\theta\mathbf{Z})$. Unfortunately, when $\mathbf{Z} = \sum_{i\in[n]}\mathbf{Z}_i$ is a random sum, bounds such as Golden-Thompson (Theorem 7) fail; extensions of Theorem 7 are false for sums of even three matrices. The matrix MGF method of [AW02], refined in [Tro11, Tro15], instead uses the following deep result.

**Proposition 4.** *Define the* quantum KL divergence *between $\mathbf{M}, \mathbf{N} \in \mathbb{S}^{d\times d}_{\succ\mathbf{0}}$ by*

$$D_H(\mathbf{N}\|\mathbf{M}) := \operatorname{Tr}\left(\mathbf{N}\left(\log\mathbf{N} - \log\mathbf{M}\right) + \mathbf{M} - \mathbf{N}\right).$$

*Then $D_H : \mathbb{S}^{d\times d}_{\succ\mathbf{0}} \times \mathbb{S}^{d\times d}_{\succ\mathbf{0}} \to \mathbb{R}$ is a jointly convex function of its inputs.*

We remark that $D_H$ in Proposition 4 is the Bregman divergence (Definition 3, Part III) of two matrices with respect to the convex function $H(\mathbf{M}) = \langle\mathbf{M}, \log(\mathbf{M})\rangle$, where convexity follows from Theorem 8 and convexity of the vector function $h(\mathbf{v}) = \langle\mathbf{v}, \log(\mathbf{v})\rangle$. Proposition 4 is perhaps not extremely surprising, as we previously gave the vector analog of this statement (Remark 5, Part III). However, the proof of Proposition 4 is sophisticated, and expands on several ideas from Section 2. For brevity we defer the proof of Proposition 4 to Chapter 4.3, [Bha07]. We state here an important consequence of it due to [Lie73], crucially used in the matrix MGF method.

**Theorem 10** (Lieb's concavity theorem)**.** *For $\mathbf{S} \in \mathbb{S}^{d\times d}$, the following is concave on $\mathbb{S}^{d\times d}_{\succ\mathbf{0}}$:*

$$f(\mathbf{X}) = \operatorname{Tr}\exp(\mathbf{S} + \log(\mathbf{X})).$$

[18] Due to the overloading of $\lambda$ with eigenvalues, in this section we use $\theta$ to parameterize the MGF.

*Proof.* First, we derive the expression in the display through a sequence of equivalences:

$$\begin{aligned}
0 &= \min_{\mathbf{N} \in \mathbb{S}^{d\times d}_{\succ \mathbf{0}}} \{D_H(\mathbf{N}\|\mathbf{M})\} \\
&= \operatorname{Tr}(\mathbf{M}) + \min_{\mathbf{N} \in \mathbb{S}^{d\times d}_{\succ \mathbf{0}}} \{\mathbf{N}(\log \mathbf{N} - \log \mathbf{M}) - \mathbf{N}\}, \\
\operatorname{Tr}(\mathbf{M}) &= \max_{\mathbf{N} \in \mathbb{S}^{d\times d}_{\succ \mathbf{0}}} \{\operatorname{Tr}(\mathbf{N}(\log \mathbf{M} - \log \mathbf{N}) + \mathbf{N})\} \\
\operatorname{Tr}\exp(\mathbf{S} + \log(\mathbf{X})) &= \max_{\mathbf{N} \in \mathbb{S}^{d\times d}_{\succ \mathbf{0}}} \{\operatorname{Tr}(\mathbf{N}(\mathbf{S} + \mathbf{I}) + \mathbf{N}(\log \mathbf{X} - \log \mathbf{N}))\} \\
&= \max_{\mathbf{N} \in \mathbb{S}^{d\times d}_{\succ \mathbf{0}}} \{\operatorname{Tr}(\mathbf{NS}) + \operatorname{Tr}(\mathbf{X}) - D_H(\mathbf{N}\|\mathbf{X})\}.
\end{aligned}$$

The first equation used that the Bregman divergence is nonnegative everywhere and the minimum $D_H(\mathbf{N}\|\mathbf{M}) = 0$ is achieved at $\mathbf{N} = \mathbf{M}$. We then applied the definition of $D_H$, and substituted $\mathbf{M} = \exp(\mathbf{S} + \log(\mathbf{X}))$. Finally, we observe that if $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a jointly convex function on its inputs, $g : \mathcal{X} \to \mathbb{R}$ defined by $g(\mathbf{x}) = \min_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y})$ is also convex when the minimum is always achieved: for $\mathbf{y} = \operatorname{argmin}_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y})$ and $\mathbf{y}' = \operatorname{argmin}_{\mathbf{y}'\in\mathcal{Y}} f(\mathbf{x}', \mathbf{y}')$, we have

$$g((1-\lambda)\mathbf{x} + \lambda\mathbf{x}') \le f((1-\lambda)\mathbf{x} + \lambda\mathbf{x}', (1-\lambda)\mathbf{y} + \lambda\mathbf{y}') \le (1-\lambda)g(\mathbf{x}) + \lambda g(\mathbf{x}').$$

Since $D_H(\mathbf{N}\|\mathbf{X}) - \operatorname{Tr}(\mathbf{NS}) - \operatorname{Tr}(\mathbf{X})$ is the sum of a jointly convex function (by Proposition 4) and a linear function, it remains jointly convex as a function of $(\mathbf{X}, \mathbf{N})$. As we argued, minimizing over $\mathbf{N}$ then yields a convex function, and negating gives the desired concavity claim. □

By repeatedly applying Theorem 10, we finally arrive at a bound compatible with Lemma 8.

**Corollary 7.** *Let $\{\mathbf{Z}_i\}_{i\in[n]} \in \mathbb{S}^{d\times d}$ be independent random matrices and $\theta \in \mathbb{R}$. Then*

$$\mathbb{E}\left[\operatorname{Tr}\exp\left(\theta \sum_{i\in[n]} \mathbf{Z}_i\right)\right] \le \operatorname{Tr}\exp\left(\sum_{i\in[n]} \log \mathbb{E}\left[\exp\left(\theta\mathbf{Z}_i\right)\right]\right).$$

*Proof.* Let $\mathbf{S}_k := \sum_{i\in[k]} \mathbf{Z}_i$ for all $k \in [n]$. We derive that

$$\begin{aligned}
\mathbb{E}\operatorname{Tr}\exp(\theta\mathbf{S}_n) &= \mathbb{E}\left[\mathbb{E}\left[\operatorname{Tr}\exp\left(\theta(\mathbf{S}_{n-1} + \mathbf{Z}_n)\right) \mid \mathbf{S}_{n-1}\right]\right] \\
&\le \mathbb{E}\left[\operatorname{Tr}\exp\left(\theta\mathbf{S}_{n-1} + \log \mathbb{E}\left[\exp(\theta\mathbf{Z}_n)\right]\right)\right] \\
&\le \mathbb{E}\left[\operatorname{Tr}\exp\left(\theta\mathbf{S}_{n-2} + \log \mathbb{E}\left[\exp(\theta\mathbf{Z}_{n-1})\right] + \log \mathbb{E}\left[\exp(\theta\mathbf{Z}_n)\right]\right)\right] \le \dots,
\end{aligned}$$

where the first line is the law of iterated expectations, and the second wrote $\theta\mathbf{Z}_n = \log\exp(\theta\mathbf{Z}_n)$ after conditioning on a realization of $\mathbf{S}_{n-1}$, and applied Jensen's inequality with Theorem 10. Finally, the last used that because of independence,[19] $\log \mathbb{E}[\exp(\mathbf{Z}_n)]$ can be treated as a deterministic matrix, and hence lumped into the sum with $\mathbf{S}_{n-2}$ when taking expectations over $\mathbf{Z}_{n-1}$. Iterating this argument to lift each $\mathbf{Z}_i$ into the exponent in sequence yields the conclusion. □

At this point, our matrix MGF strategy is clear. Because $\operatorname{Tr}\exp$ is a monotone function (Corollary 6), to bound the right-hand side in Corollary 7, it suffices to provide upper bounds on each of the matrices $\log \mathbb{E}\left[\exp\left(\theta\mathbf{Z}_i\right)\right]$, which can be controlled eigenvalue-by-eigenvalue using moments. We can then substitute this bound into Lemma 8 to obtain our desired tail bounds.

## 3.2 Matrix concentration inequalities

In this section, we first give some bounds on $\log \mathbb{E}[\exp(\theta\mathbf{Z})]$, where $\mathbf{Z}$ is a random matrix satisfying various structural assumptions. We then combine these bounds with the matrix MGF framework in Lemma 8 and Corollary 7 to derive our main matrix concentration inequalities.

We mention that we will only extend Theorem 2 (in the special case of bounded random variables, i.e., Lemma 2) and Theorem 3 to the matrix setting, neglecting Theorem 1. This is in part because

[19] This argument can be extended to the matrix martingale setting (see Remark 2) as shown in Section 7, [Tro11], via a symmetrization trick, but this requires some more effort so we omit it.

some of the proof techniques used to establish sub-Gaussianity of random variables (e.g., Lemma 1) do not extend to matrices, because exp is not operator convex. However, in several important cases (such as matrix Rademacher and matrix Gaussian series) one can directly bound the matrix MGF and generalize Theorem 1; we defer more details to Chapter 4 of [Tro15].

We now give a matrix variant of Lemma 2, used to control the MGF of bounded random variables.

**Lemma 9.** *Let $\mathbf{Z} \in \mathbb{S}^{d\times d}$ be a random matrix satisfying $\mathbb{E}\mathbf{Z} = \mathbf{0}_d$ and $\|\mathbf{Z}\|_{\mathrm{op}} \le c$ with probability 1. Then, for all $|\theta| \le \frac{1}{2c}$, we have $\log \mathbb{E}\left[\exp\left(\theta\mathbf{Z}\right)\right] \preceq \theta^2\mathbb{E}[\mathbf{Z}^2]$.*

*Proof.* For a fixed realization of $\mathbf{Z}$ with eigendecomposition $\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, we have

$$\exp\left(\theta\mathbf{Z}\right) = \mathbf{I}_d + \theta\mathbf{Z} + \sum_{p=2}^{\infty}\frac{\theta^p}{p!}\mathbf{Z}^p \preceq \mathbf{I}_d + \theta\mathbf{Z} + \sum_{p=2}^{\infty}\frac{\theta^p c^{p-2}}{p!}\mathbf{Z}^2 \preceq \mathbf{I}_d + \theta\mathbf{Z} + \theta^2\mathbf{Z}^2,$$

where we used that for $p \ge 2$, $\mathbf{Z}^p \preceq c^{p-2}\mathbf{Z}^2$ by Lemma 6 and $x^p \le c^{p-2}x^2$ for $|x| \le c$. We then applied (3). The claim follows by taking expectations (since $\mathbb{E}\mathbf{Z} = \mathbf{0}_d$), applying $\log(1+x) \le x$ for all $x \in \mathbb{R}$, and again lifting this scalar inequality to commuting matrices via Lemma 6. ☐

The matrix Bernstein inequality then follows readily from Lemma 9 and our earlier framework.

**Theorem 11** (Matrix Bernstein's inequality)**.** *Let $\{\mathbf{Z}_i\}_{i\in[n]} \in \mathbb{S}^{d\times d}$ be independent random matrices satisfying $\mathbb{E}\mathbf{Z}_i = \mathbf{0}_d$ and $\|\mathbf{Z}_i\|_{\mathrm{op}} \le c$ with probability 1. Further, let $\mathbf{Z} := \sum_{i\in[n]}\mathbf{Z}_i$ and suppose $\mathbb{E}\sum_{i\in[n]}\mathbf{Z}_i^2 \preceq \sigma^2\mathbf{I}_d$. Then for all $t \ge 0$,*

$$\Pr\left[\|\mathbf{Z}\|_{\mathrm{op}} \ge t\right] \le 2d\exp\left(-\min\left(\frac{t^2}{4\sigma^2}, \frac{t}{4c}\right)\right).$$

*Proof.* By combining Corollary 7 and Lemma 9, we have for all $\theta \in [0, \frac{1}{2c}]$,

$$\mathbb{E}\left[\mathrm{Tr}\exp\left(\theta\mathbf{Z}\right)\right] \le \mathrm{Tr}\exp\left(\theta^2\mathbb{E}\left[\sum_{i\in[n]}\mathbf{Z}_i^2\right]\right) \le \mathrm{Tr}\exp\left(\theta^2\sigma^2\mathbf{I}_d\right) = d\exp(\theta^2\sigma^2),$$

where we used monotonicity of $\mathrm{Tr}\exp$ (Corollary 6) and the definition of $\sigma$. The conclusion follows from applying the upper tail bound in Lemma 8 with $\theta = \min(\frac{t}{2\sigma^2}, \frac{1}{2c})$ (as in Theorem 2), and noticing that for the lower tail bound it suffices to negate all matrices by symmetry. ☐

Theorem 11 gives exactly the same bound as what one would attain in the scalar case (when $d = 1$) by combining Lemma 2, Fact 3, and Theorem 2. Generally, the bound decays by a $d$ factor, which roughly corresponds to the loss due to a union bound because we must control $d$ eigenvalues simultaneously. We also make the following remark on lifting the symmetric matrix assumption.

**Remark 4.** *Theorem 11 extends to sums of asymmetric random matrices via a tiling argument. Specifically, suppose that $\{\mathbf{A}_i\}_{i\in[n]} \in \mathbb{R}^{d_1\times d_2}$ are independent random matrices satisfying $\mathbb{E}\mathbf{A}_i = \mathbf{0}_{d_1\times d_2}$ and $\|\mathbf{A}_i\|_{\mathrm{op}} \le c$ for all $i \in [n]$. Then, let $d = d_1 + d_2$, and for each $i \in [n]$ define*

$$\mathbf{Z}_i := \begin{pmatrix}\mathbf{0}_{d_1} & \mathbf{A}_i\\ \mathbf{A}_i^\top & \mathbf{0}_{d_2}\end{pmatrix}.$$

*It is simple to check that $\|\mathbf{Z}_i\|_{\mathrm{op}} = \|\mathbf{A}_i\|_{\mathrm{op}}$, $\mathbb{E}\mathbf{Z}_i = \mathbf{0}_d$, and*

$$\mathbb{E}\left[\sum_{i\in[n]}\mathbf{Z}_i^2\right] = \begin{pmatrix}\mathbb{E}\left[\sum_{i\in[n]}\mathbf{A}_i\mathbf{A}_i^\top\right] & \mathbf{0}_{d_1\times d_2}\\ \mathbf{0}_{d_2\times d_1} & \mathbb{E}\left[\sum_{i\in[n]}\mathbf{A}_i^\top\mathbf{A}_i\right]\end{pmatrix}.$$

*Thus, Theorem 11 applies and depends on $\sigma^2 \ge \max(\boldsymbol{\lambda}_1(\mathbb{E}\sum_{i\in[n]}\mathbf{A}_i\mathbf{A}_i^\top), \boldsymbol{\lambda}_1(\mathbb{E}\sum_{i\in[n]}\mathbf{A}_i^\top\mathbf{A}_i))$.*

Next, we give an inequality which applies to random positive semidefinite matrices.

**Lemma 10.** *Let $\mathbf{Z} \in \mathbb{S}^{d\times d}_{\succeq\mathbf{0}}$ be a random matrix satisfying $\boldsymbol{\lambda}_1(\mathbf{Z}) \le R$ with probability 1. Then for all $\theta \in \mathbb{R}$, we have $\log\mathbb{E}\left[\exp(\theta\mathbf{Z})\right] \preceq \frac{\exp(\theta R)-1}{R}\mathbb{E}[\mathbf{Z}]$.*

*Proof.* Because $\exp(\theta x)$ is convex for $x \in \mathbb{R}$, we have $\exp(\theta x) \le 1 + \frac{\exp(\theta R)-1}{R} x$ for all $x \in [0, R]$, by considering the line from $(0, 1)$ to $(R, \exp(\theta R))$, which the point $(x, \exp(\theta x))$ lies below. The conclusion follows from Lemma 6, since $\log(1 + t) \le t$ for all $t \in \mathbb{R}$. □

We conclude by plugging in Lemma 10 into our matrix MGF method, to extend Theorem 3.

**Theorem 12** (Matrix Chernoff bound)**.** *Let* $\{\mathbf{Z}_i\}_{i\in[n]} \in \mathbb{S}^{d\times d}$ *be independent random matrices satisfying* $\|\mathbf{Z}_i\|_{\mathrm{op}} \le R$ *with probability* 1*, and let* $\mathbf{Z} := \sum_{i\in[n]} \mathbf{Z}_i$*. Then,*

$$\Pr\left[\lambda_1(\mathbf{Z}) \ge (1+\epsilon)\mu_{\max}\right] \le d \exp\left(-\frac{\epsilon^2 \mu_{\max}}{(2+\epsilon)R}\right) \text{ for all } \epsilon \ge 0, \text{ where } \mu_{\max} := \lambda_1(\mathbb{E}\mathbf{Z}),$$
$$\Pr\left[\lambda_d(\mathbf{Z}) \le (1-\epsilon)\mu_{\min}\right] \le d \exp\left(-\frac{\epsilon^2 \mu_{\min}}{2R}\right) \text{ for all } \epsilon \in (0,1), \text{ where } \mu_{\min} := \lambda_d(\mathbb{E}\mathbf{Z}).$$

*Proof.* Let $f(\theta) := \frac{\exp(\theta R)-1}{R}$. We begin with the tail bound on $\lambda_1(\mathbf{Z})$. For all $\theta \ge 0$, combining Corollary 7 and Lemma 10, and using the definition of $\mu_{\max}$, shows that

$$\mathbb{E}\left[\operatorname{Tr}\exp\left(\theta\mathbf{Z}\right)\right] \le \operatorname{Tr}\exp\left(f(\theta)\sum_{i\in[n]}\mathbb{E}[\mathbf{Z}_i]\right) = \operatorname{Tr}\exp\left(f(\theta)\mathbb{E}\mathbf{Z}\right) \le d\exp(f(\theta)\mu_{\max}).$$

Therefore, by the upper tail bound in Lemma 8,

$$\Pr\left[\lambda_1(\mathbf{Z}) \ge (1+\epsilon)\mu_{\max}\right] \le \inf_{\theta>0} d\exp\left(\left(f(\theta) - \theta(1+\epsilon)\right)\mu_{\max}\right) \le d\exp\left(-\frac{\epsilon^2\mu_{\max}}{(2+\epsilon)R}\right),$$

where we plugged in $\theta \leftarrow \frac{1}{R}\log(1+\epsilon)$, and used the estimate $\log(1+\epsilon) \ge \frac{2\epsilon}{2+\epsilon}$ for $\epsilon \ge 0$. Similarly, for the tail bound on $\lambda_d(\mathbf{Z})$, we have for all $\theta \le 0$ that $f(\theta) \le 0$, so

$$\mathbb{E}[\operatorname{Tr}\exp(\theta\mathbf{Z})] \le \operatorname{Tr}\exp\left(f(\theta)\mathbb{E}\mathbf{Z}\right) \le d\exp\left(f(\theta)\mu_{\min}\right).$$

Therefore, by the lower tail bound in Lemma 8,

$$\Pr\left[\lambda_d(\mathbf{Z}) \le (1-\epsilon)\mu_{\min}\right] \le \inf_{\theta<0} d\exp\left(\left(f(\theta) - \theta(1-\epsilon)\right)\mu_{\min}\right) \le d\exp\left(-\frac{\epsilon^2\mu_{\min}}{2R}\right),$$

where the last inequality chose $\theta \leftarrow \frac{1}{R}\log(1-\epsilon)$ and used $\frac{\epsilon^2}{2} \le \epsilon + (1-\epsilon)\log(1-\epsilon)$ for $\epsilon \in (0,1)$. □

## Source material

Portions of this lecture are based on reference material in [Bha07, BLM13, Tro15, vH16, Ver18, Lee21, Duc23], as well as the author's own experience working in the field.

## References


[AW02] R. Ahlswede and A. Winter. Strong converse for identification via quantum channels. *IEEE Trans. Inform. Theory*, 48(3):569–579, 2002.

[Bha07] Rajendra Bhatia. *Positive Definite Matrices*. Princeton University Press, 2007.

[BLM13] Stéphane Boucheron, Gabor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[Duc23] John Duchi. *Lecture Notes on Statistics and Information Theory*. 2023.

[Eld13] Ronen Eldan. Thin shell implies spectral gap via a stochastic localization scheme. *Geom. Funct. Anal.*, 23(2):532–569, 2013.

[Gol65] Sidney Golden. Lower bounds for the helmholtz function. *Phys. Rev., Series II*, 137(4B):B1127–B1128, 1965.

[Lee21] James Lee. *The art and science of positive definite matrices*. 2021.

[Lew96] Adrian Lewis. Convex analysis on the hermitian matrices. *SIAM Journal on Optimization*, 6(0):164–177, 1996.

[Lie73] E. H. Lieb. Convex trace functions and the wigner-yanase-dyson conjecture. *Adv. Math.*, 11:267–288, 1973.

[LM00] Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

[LS01] Adrian Lewis and Hristo S. Sendov. Twice differentiable spectral functions. *SIAM Journal on Matrix Analysis and Applications*, 23(0):368–386, 2001.

[LT76] E. H. Lieb and W. E. Thirring. Inequalities for the moments of the eigenvalues of the schrödinger hamiltonian and their relation to sobolev inequalities. *Studies in Mathematical Physics*, pages 269–303, 1976.

[Nes07] Y. Nesterov. Smoothing technique and its applications in semidefinite optimization. *Math. Program.*, 110:245–259, 2007.

[Tao10] Terence Tao. The golden-thompson inequality. https://terrytao.wordpress.com/2010/07/15/the-golden-thompson-inequality/, 2010. Accessed: 2024-01-01.

[Tho65] Colin J. Thompson. Inequality with applications in statistical mechanics. *Journal of Mathematical Physics*, 6(12):1812–1813, 1965.

[Tro11] Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12:389–434, 2011.

[Tro15] Joel A. Tropp. An introduction to matrix concentration inequalities. *Found. Trends Mach. Learn.*, 8(1-2):1–230, 2015.

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

[vH16] Ramon van Handel. *Probability in High Dimension*. 2016.

[vN37] John von Neumann. Some matrix-inequalities and metrization of matrix-space. *Tomsk Univ. Rev. 1*, pages 286–300, 1937.

[ZLO16] Zeyuan Allen Zhu, Yin Tat Lee, and Lorenzo Orecchia. Using optimization to obtain a width-independent, parallel, simpler, and faster positive SDP solver. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016*, pages 1824–1831. SIAM, 2016.