**Lecture 12: Concentration for sums of random matrices** CSE 525, Spring 2019
**Instructor:** James R. Lee

# 1 Random matrices

Consider the *graph sparsification problem:* Given a graph $G = (V, E)$, we want to approximate $G$ (in a sense to be defined later) by a sparse graph $H = (V, E')$. Generally we would like that $E' \subseteq E$ and moreover $|E'|$ is as small as possible—say $O(n)$ or $O(n \log n)$ where $n = |V|$. We will be able to do this by choosing a (nonuniform) random sample of the edges, but to analyze such a process, we will need a large-deviation inequality for sums of random *matrices.*

## 1.1 Symmetric matrices

If $A$ is a $d \times d$ real symmetric matrix, then $A$ has all real eigenvalues which we can order $\lambda_1(A) \geqslant \lambda_2(A) \geqslant \cdots \geqslant \lambda_d(A)$. The *operator norm* of $A$ is

$$\|A\| := \max_{\|x\|_2=1} \|Ax\|_2 = \max\{|\lambda_i(A)| : i \in \{1, \ldots, d\}\} .$$

The trace of $A$ is $\mathrm{Tr}(A) = \sum_{i=1}^d A_{ii} = \sum_{i=1}^d \lambda_i(A)$. The *trace norm* of $A$ is $\|A\|_* = \sum_{i=1}^d |\lambda_i(A)|$. A symmetric matrix is *positive semidefinite* (PSD) if all its eigenvalues are nonnegative. Note that for a PSD matrix $A$, we have $\mathrm{Tr}(A) = \|A\|_*$. We also recall the matrix exponential $e^A = \sum_{k=0}^\infty \frac{A^k}{k!}$ which is well-defined for all real symmetric $A$ and is itself also a real symmetric matrix. If $A$ is symmetric, then $e^A$ is always PSD, as the next argument shows.

Every real symmetric matrix can be diagonalized, writing $A = U^T D U$, where $U$ is an orthogonal matrix, i.e. $UU^T = U^T U = I$ and $D$ is diagonal. One can easily check that $A^k = U^T D^k U$ for any $k \in \mathbb{N}$, thus $A^k$ and $A$ are simultaneously diagonalizable. It follows that $A$ and $e^A$ are simultaneously diagonalizable. In particular, we have $\lambda_i(e^A) = e^{\lambda_i(A)}$.

Finally, note that for symmetric matrices $A$ and $B$, we have $|\mathrm{Tr}(AB)| \leqslant \|A\| \cdot \|B\|_*$. To see this, let $\{u_i\}$ be an orthonormal basis of eigenvectors of $B$ with $Bu_i = \lambda_i(B)u_i$. Then

$$|\mathrm{Tr}(AB)| = \left|\sum_{i=1}^d \langle u_i, ABu_i\rangle\right| = \left|\sum_{i=1}^d \lambda_i(B)\langle u_i, Au_i\rangle\right| \leqslant \sum_{i=1}^d |\lambda_i(B)| \cdot \|A\| = \|B\|_*\|A\|.$$

Many classical statements are either false or significantly more difficult to prove when translated to the matrix setting. For instance, while $e^{x+y} = e^x e^y = e^y e^x$ is true for arbitrary real numbers $x$ and $y$, it is only the case that $e^{A+B} = e^A e^B$ if $A$ and $B$ are simultaneously diagonalizable. However, somewhat remarkably, the matrix analog does hold if we do it inside the trace.

**Theorem 1.1** (Golden-Thompson inequality)**.** *If $A$ and $B$ are real symmetric matrices, then*

$$\mathrm{Tr}(e^{A+B}) \leqslant \mathrm{Tr}(e^A e^B) .$$

*Proof.* We can prove this using the non-commutative Hölder inequality: For any even integer $p \geqslant 2$ and real symmetric matrices $A_1, A_2, \ldots, A_p$:

$$\left|\mathrm{Tr}(A_1 A_2 \cdots A_p)\right| \leqslant \|A_1\|_{S_p}\|A_2\|_{S_p} \cdots \|A_p\|_{S_p},$$

where $\|A\|_{S_p} = \left(\mathrm{Tr}(A^T A)^{p/2}\right)^{1/p}$ is the Schatten $p$-norm. Consider real symmetric matrices $U, V$. Applying this with $A_1 = \cdots A_p = UV$ gives, for every even $p \geqslant 2$:

$$\mathrm{Tr}((UV)^p) \leqslant \|UV\|_{S_p}^p = \mathrm{Tr}((V^T U^T UV)^{p/2}) = \mathrm{Tr}((VU^2V)^{p/2}) = \mathrm{Tr}((U^2V^2)^{p/2}),$$

where the last inequality uses the cyclic property of the trace. Applying this inequality repeatedly now yields, for every even $p$,

$$\mathrm{Tr}((UV)^p) \leqslant \mathrm{Tr}(U^p V^p).$$

If we now take $U = e^{A/p}$ and $V = e^{B/p}$, this gives

$$\mathrm{Tr}((e^{A/p} e^{B/p})^p) \leqslant \mathrm{Tr}(e^A e^B). \tag{1.1}$$

For $p$ large, we can use the Taylor approximation $e^{A/p} = 1 + A/p + O(1/p^2)$ and similarly for $e^{B/p}$. Thus: $e^{A/p} e^{B/p} \sim e^{(A+B)/p + O(1/p^2)}$. Therefore taking $p \to \infty$ in (1.1) gives

$$\mathrm{Tr}(e^{A+B}) \leqslant \mathrm{Tr}(e^A e^B). \qquad \square$$

## 1.2 The Laplace transform for matrices

We will consider now a random $d \times d$ real matrix $X$. The entries $(X_{ij})$ of $X$ are all (not necessarily independent) random variables. We have seen inequalities (like those named after Chernoff and Azuma) which assert that if $X = X_1 + X_2 + \cdots + X_n$ is a sum of *independent* random numbers, then $X$ is tightly concentrated around its mean. Our goal now is to prove a similar fact for sums of independent random symmetric matrices.

First, observe that the trace is a linear operator; this is easy to see from the fact that it is the sum of the diagonal entries of its argument. If $A$ and $B$ are arbitrary real matrices, then $\mathrm{Tr}(A + B) = \mathrm{Tr}(A) + \mathrm{Tr}(B)$. This implies that if $X$ is a random matrix, then $\mathbb{E}[\mathrm{Tr}(X)] = \mathrm{Tr}(\mathbb{E}[X])$. Note that $\mathbb{E}[X]$ is the matrix defined by $(\mathbb{E}[X])_{ij} = \mathbb{E}[X_{ij}]$.

Suppose that $X_1, X_2, \ldots, X_n$ are independent random real symmetric matrices. Let $X = X_1 + X_2 + \cdots + X_n$. Let $S_k = X_1 + \cdots + X_k$ be the partial sum of the first $k$ terms so that $X = S_n$. Our first goal will be to bound the probability that $X$ has an eigenvalue bigger than $t$. To do this, we will try to extend the method of exponential moments to work with symmetric matrices, as discovered by Ahlswede and Winter. It is much simpler than previous approaches that only worked for special cases.

Note that for $\beta > 0$, we have $\lambda_i(e^{\beta X}) = e^{\beta \lambda_i(X)}$. Therefore:

$$\mathbb{P}\left[\max_i \lambda_i(X) > t\right] = \mathbb{P}\left[\max_i \lambda_i(e^{\beta X}) > e^{\beta t}\right] \leqslant \mathbb{P}\left[\mathrm{Tr}(e^{\beta X}) > e^{\beta t}\right], \tag{1.2}$$

where the last inequality uses the fact that all the eigenvalues of $e^{\beta X}$ are nonnegative, hence $\mathrm{Tr}(e^{\beta X}) = \sum_i \lambda_i(e^{\beta X}) \geqslant \max_i \lambda_i(e^{\beta X})$.

Now Markov's inequality implies that

$$\mathbb{P}[\mathrm{Tr}(e^{\beta X}) > e^{\beta t}] \leqslant \frac{\mathbb{E}[\mathrm{Tr}(e^{\beta X})]}{e^{\beta t}}. \tag{1.3}$$

As in our earlier uses of the Laplace transform, our goal is now to bound $\mathbb{E}[\mathrm{Tr}(e^{\beta X})]$ by a product that has one factor for each term $X_i$.

In the matrix setting, this is more subtle: Using Theorem 1.1,

$$\mathbb{E}[\mathrm{Tr}(e^{\beta X})] = \mathbb{E}[\mathrm{Tr}(e^{\beta(S_{n-1}+X_n)})] \leqslant \mathbb{E}[\mathrm{Tr}(e^{\beta S_{n-1}} e^{\beta X_n})] .$$

Now we push the expectation over $X_n$ inside the trace:

$$\mathbb{E}[\mathrm{Tr}(e^{\beta S_{n-1}} e^{\beta X_n})] = \mathbb{E}\left[\mathrm{Tr}(e^{\beta S_{n-1}} \mathbb{E}[e^{\beta X_n} \mid X_1, \ldots, X_{n-1}])\right] = \mathbb{E}\left[\mathrm{Tr}(e^{\beta S_{n-1}} \mathbb{E}[e^{\beta X_n}])\right] ,$$

and we have used independence to pull $e^{\beta S_{n-1}}$ outside the expectation and then to remove the conditioning. Finally, we use the fact that $\mathrm{Tr}(AB) \leqslant \|A\| \cdot \|B\|_*$ and $\|B\|_* = \mathrm{Tr}(B)$ when $B$ has all nonnegative eigenvalues (as is the case for $e^{\beta S_{n-1}}$):

$$\mathbb{E}\left[\mathrm{Tr}(e^{\beta S_{n-1}} \mathbb{E}[e^{\beta X_n}])\right] \leqslant \left\|\mathbb{E}[e^{\beta X_n}]\right\| \mathbb{E}\left[\mathrm{Tr}(e^{\beta S_{n-1}})\right] .$$

Doing this $n$ times yields

$$\mathbb{E}[\mathrm{Tr}(e^{\beta X})] \leqslant \mathrm{Tr}(I) \prod_{i=1}^{n} \left\|\mathbb{E}[e^{\beta X_i}]\right\| = d \prod_{i=1}^{n} \left\|\mathbb{E}[e^{\beta X_i}]\right\| .$$

Combining this with (1.2) and (1.3) yields

$$\mathbb{P}\left[\max_i \lambda_i(X) > t\right] \leqslant e^{-\beta t} d \prod_{i=1}^{n} \left\|\mathbb{E}[e^{\beta X_i}]\right\| .$$

We can also apply this to $-X$ to get

$$\mathbb{P}\left[\|X\| > t\right] \leqslant e^{-\beta t} d \left(\prod_{i=1}^{n} \left\|\mathbb{E}[e^{\beta X_i}]\right\| + \prod_{i=1}^{n} \left\|\mathbb{E}[e^{-\beta X_i}]\right\|\right) . \tag{1.4}$$

### 1.3 Concentration

Let $Y$ be a random, symmetric, psd $d \times d$ matrix with $\mathbb{E}[Y] = I$. Suppose that $\|Y\| \leqslant L$ with probability one.

**Theorem 1.2.** *If $Y_1, Y_2, \ldots, Y_n$ are i.i.d. copies of $Y$, then for any $\varepsilon \in (0,1)$ the following holds. Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ denote the eigenvalues of $\frac{1}{n}\sum_{i=1}^{n} Y_i$. Then*

$$\mathbb{P}[\{\lambda_1, \lambda_2, \ldots, \lambda_n\} \subseteq [1-\varepsilon, 1+\varepsilon]] \geqslant 1 - 2d \exp\left(-\varepsilon^2 n/4L\right) .$$

There is a slightly nicer way to write this using the *Löwner ordering of symmetric matrices:* We write $A \succeq B$ to denote that the matrix $A - B$ is positive semidefinite. We can rewrite the conclusion of Theorem 1.2 as

$$\mathbb{P}\left[(1-\varepsilon)I \preceq \frac{1}{n}\sum_{i=1}^{n} Y_i \preceq (1+\varepsilon)I\right] \geqslant 1 - 2d \exp\left(-\varepsilon^2 n/4L\right) . \tag{1.5}$$

*Proof of Theorem 1.2.* Define $X_i := Y_i - \mathbb{E}[Y_i]$ and $X := X_1 + \cdots + X_n$. Then (1.5) is equivalent to

$$\mathbb{P}\left[\|X\| > \varepsilon n\right] \leqslant 2d \exp\left(-\varepsilon^2 n/4L\right) .$$

We know from (1.4) that it will suffice to bound $\left\|\mathbb{E}[e^{\beta X_i}]\right\|$ for each $i$. To do this, we will use the fact that

$$1 + x \leqslant e^x \leqslant 1 + x + x^2 \qquad \forall x \in [-1, 1] .$$

Note that if $A$ is a real symmetric matrix, then since $I, A, A^2$, and $e^A$ are simultaneously diagonalizable, this yields

$$I + A \preceq e^A \preceq I + A + A^2 \tag{1.6}$$

for any $A$ with $\|A\| \leqslant 1$.

Observe that $\mathbb{E}[X_i] = 0$. To evaluate $\|X_i\|$, let us use the fact that for real symmetric $A$, we have $\|A\| = \max_{\|x\|_2=1} |x^T A x|$. So consider some $x \in \mathbb{R}^d$ with $\|x\|_2 = 1$ and write

$$|x^T X_i x| = \frac{|x^T Y_i x - x^T \mathbb{E}[Y_i] x|}{n} \leqslant |x^T Y_i x| \leqslant L\,,$$

where we have used the fact that since $Y_i$ is PSD, so is $\mathbb{E}[Y_i]$, and thus $x^T \mathbb{E}[Y_i] x$ and $x^T Y_i x$ are both nonnegative. We also used our assumption that $\|Y_i\| \leqslant L$. We conclude that $\|X_i\| \leqslant L$.

Moreover, we have

$$\mathbb{E}[X_i^2] = \mathbb{E}[(Y_i - \mathbb{E}[Y_i])^2] = \left(\mathbb{E}[Y_i^2] - (\mathbb{E}[Y_i])^2\right) \preceq \mathbb{E}[Y_i^2] \preceq \mathbb{E}[\|Y_i\| Y_i] \preceq L\,\mathbb{E}[Y_i]\,,$$

where in the final line we again used the assumption $\|Y_i\| \leqslant L$. Finally, since $\mathbb{E}[Y_i] = I$, conclude that $\|\mathbb{E}[X_i^2]\| \leqslant L$.

Therefore for any $\beta \leqslant 1/L$, we can apply (1.6), yielding

$$\mathbb{E}[e^{\beta X_i}] \preceq I + \beta\,\mathbb{E}[X_i] + \beta^2\,\mathbb{E}[X_i^2] = I + \beta^2\,\mathbb{E}[X_i^2] \preceq e^{\beta^2\,\mathbb{E}[X_i^2]}\,.$$

We conclude that for $\beta \leqslant 1/L$, we have $\|\mathbb{E}[e^{\beta X_i}]\| \leqslant e^{\beta^2 L}$.

Plugging this into (1.4), we see that

$$\mathbb{P}\left[\|X\| > \varepsilon n\right] \leqslant 2de^{-\varepsilon n\beta} e^{\beta^2 L n}.$$

Choosing $\beta := \frac{\varepsilon}{2L}$ yields

$$\mathbb{P}\left[\|X\| > \varepsilon n\right] \leqslant 2de^{-\varepsilon^2 n/4L}\,,$$

completing the argument. □

Finally, we can prove a generalization of Theorem 1.2 for random matrices whose expectation is not the identity.

**Theorem 1.3.** *Let $Z$ be a $d \times d$ random real, symmetric, PSD matrix. Suppose also that $Z \preceq L \cdot \mathbb{E}[Z]$ for some $L \geqslant 1$. If $Z_1, Z_2, \ldots, Z_n$ are i.i.d. copies of $Z$, then for any $\varepsilon > 0$, it holds that*

$$\mathbb{P}\left[(1-\varepsilon)\,\mathbb{E}[Z] \preceq \frac{1}{n}\sum_{i=1}^n Z_i \preceq (1+\varepsilon)\,\mathbb{E}[Z]\right] \geqslant 1 - 2de^{-\varepsilon^2 n/4L}\,.$$

In other words, the empirical mean $\frac{1}{n}\sum_{i=1}^n Z_i$ is very close (in a spectral sense) to the expectation $\mathbb{E}[Z]$.

*Proof of Theorem 1.3.* This is made difficult only because it may be that $A = \mathbb{E}[Z]$ is not invertible. Suppose that $A = UDU^T$ where $D$ is a diagonal matrix $D = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k, 0, \ldots, 0)$, where $k$ is the rank of $A$ and $\lambda_i \neq 0$ for $i = 1, \ldots, k$. Then the *pseudoinverse* of $A$ is defined by

$$A^+ = U\mathrm{diag}(\lambda_1^{-1}, \lambda_2^{-1}, \ldots, \lambda_k^{-1}, 0, \ldots, 0)U^T\,.$$

Note that $AA^+ = A^+A = I_{\mathrm{im}(A)}$, where $I_{\mathrm{im}(A)}$ denotes the operator that acts by the identity on the image of $A$, and annihilates $\ker(A)$.

Since $A$ is PSD, $A^+$ is also PSD, and we can define $A^{+/2}$ as the *square root* of $A^+$. One can write this explicitly as

$$A^{+/2} = U\mathrm{diag}(\lambda_1^{-1/2}, \lambda_2^{-1/2}, \ldots, \lambda_k^{-1/2}, 0, \ldots, 0)U^T .$$

Now to prove Theorem 1.3, it suffices to apply Theorem 1.2 to the matrices $Y_i = A^{+/2}Z_iA^{+/2}$. Verification is left as an exercise. □