Lecture 2: Sums of independent random matrices CSE 599I, Spring 2021
Instructor: James R. Lee

1 Concentration inequalities

The style of this course will be to consider some phenomenon with real numbers and then to explore what happens when we try to translate it to the world of symmetric matrices. Our first example concerns tail bounds for sums of independent random variables.

The classical setting. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be independent random variables and denote $\boldsymbol{X} := \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n$. Let us additionally suppose that $|\boldsymbol{X}_i - \mathbb{E}\,\boldsymbol{X}_i| \leqslant L$ holds almost surely for every $i = 1, 2, \ldots, n$. Then we have the classical exponential concentration inequality: For every $t \geqslant 0$,

$$\mathbb{P}\left[\boldsymbol{X} \geqslant \mathbb{E}\,\boldsymbol{X} + t\sqrt{n}\right] \leqslant \exp\left(\frac{-t^2}{2L^2}\right). \tag{1.1}$$

By translation, we may assume that $\mathbb{E}\,\boldsymbol{X}_i = 0$ and $|\boldsymbol{X}_i| \leqslant L$ for every i. The standard argument proceeds by applying Markov's inequality to the Laplace transform $\beta \mapsto e^{\beta \boldsymbol{X}}$:

$$\mathbb{P}\left[\boldsymbol{X} \geqslant t\sqrt{n}\right] = \mathbb{P}\left[e^{\beta \boldsymbol{X}} \geqslant e^{\beta t\sqrt{n}}\right] \leqslant e^{-\beta t\sqrt{n}}\,\mathbb{E}\left[e^{\beta \boldsymbol{X}}\right], \tag{1.2}$$

where $\beta > 0$ is some parameter we will optimize over later.

The advantage of the exponential representation is that it allows us to use mutual independence:

$$\mathbb{E}\left[e^{\beta \boldsymbol{X}}\right] = \mathbb{E}\left[e^{\beta(\boldsymbol{X}_1 + \cdots \boldsymbol{X}_n)}\right] = \prod_{i=1}^{n} \mathbb{E}[e^{\beta \boldsymbol{X}_i}]. \tag{1.3}$$

Note that

$$1 + x \leqslant e^x \leqslant 1 + x + x^2, \quad \forall x \in [-1, 1]. \tag{1.4}$$

Therefore if we ensure that $\beta \leqslant 1/L$, we have

$$\mathbb{E}[e^{\beta \boldsymbol{X}_i}] \leqslant \mathbb{E}[1 + \beta \boldsymbol{X}_i + \beta^2 \boldsymbol{X}_i^2] = 1 + \beta^2\,\mathbb{E}[\boldsymbol{X}_i^2] \leqslant 1 + \beta^2 L^2 \leqslant e^{\beta^2 L^2}.$$

Plugging this into (1.2) and (1.3) gives: For $0 < \beta \leqslant 1/L$,

$$\mathbb{P}\left[\boldsymbol{X} \geqslant t\sqrt{n}\right] \leqslant e^{-\beta t\sqrt{n} + \beta^2 L^2 n}. \tag{1.5}$$

Now setting $\beta := \frac{t}{2L^2\sqrt{n}}$ gives

$$\mathbb{P}\left[\boldsymbol{X} \geqslant t\sqrt{n}\right] \leqslant \exp\left(\frac{-t^2}{2L^2}\right),$$

as long as $t \leqslant 2L\sqrt{n}$. But under our assumptions, $\boldsymbol{X} \leqslant nL$ almost surely, hence the bound is true for all $t \geqslant 0$.

Sums of independent symmetric matrices. Let us now suppose that $\boldsymbol{A}_1, \dots, \boldsymbol{A}_n \in \mathbb{M}_d$ is a family of independent random symmetric matrices and $\boldsymbol{A} := \boldsymbol{A}_1 + \cdots + \boldsymbol{A}_n$. Finding the correct analog of (1.1) might, in general, be driven by the application we have in mind.

A natural approach is to replace numbers by matrices and inequalities by the Loewner order. This would yield the assumptions: Almost surely,

$$-L \cdot I \preceq \boldsymbol{A}_i - \mathbb{E}\,\boldsymbol{A}_i \preceq L \cdot I, \quad \forall i = 1, \dots, n, \tag{1.6}$$

and the conclusion

$$\mathbb{P}\left[\boldsymbol{A} \succeq \mathbb{E}\,\boldsymbol{A} + t\sqrt{n} \cdot I\right] \leqslant \exp\left(\frac{-t^2}{2L^2}\right).$$

Actually, this conclusion is substantially weaker than what we might hope for. Indeed, let us assume that $\mathbb{E}\,\boldsymbol{A}_i = 0$ for all i. Then the "bad" event $\{\boldsymbol{A} \succeq t\sqrt{n} \cdot I\}$ entails *all* the eigenvalues of $\boldsymbol{A}$ being large, whereas we might hope to argue that *none of them* are large.

Our naive translation of (1.1) is due to the fact that $\preceq$ is not a total order: $A \succeq B$ is not the logical negation of $A \prec B$. If we rewrite (1.1) as

$$\mathbb{P}\left[\boldsymbol{X} \leqslant \mathbb{E}\,\boldsymbol{X} + t\sqrt{n}\right] \geqslant 1 - \exp\left(\frac{-t^2}{2L^2}\right),$$

then taking again the straightforward translation to matrices yields

$$\mathbb{P}\left[\boldsymbol{A} \preceq \mathbb{E}\,\boldsymbol{A} + t\sqrt{n} \cdot I\right] \geqslant 1 - \exp\left(\frac{-t^2}{2L^2}\right). \tag{1.7}$$

By translation, we may again assume that $\mathbb{E}\,\boldsymbol{A}_i = 0$ for each i, and then (1.6) can be restated as: Almost surely,

$$\|\boldsymbol{A}_i\| \leqslant L, \quad \forall i = 1, \dots, n,$$

where

$$\|M\| := \sup_{x \neq 0} \frac{\|Mx\|_2}{\|x\|_2}$$

denotes the *operator norm* of $M \in \mathbb{M}_d$.

When A is symmetric, $\|A\| = \max(|\lambda_1|, \dots, |\lambda_d|)$, where $\lambda_1, \dots, \lambda_d$ are the eigenvalues of A. And since $\mathbb{E}\,\boldsymbol{A} = 0$, the bound (1.7) can be rewritten as

$$\mathbb{P}\left[\lambda_{\max}(\boldsymbol{A}) \geqslant t\sqrt{n}\right] \leqslant \exp\left(\frac{-t^2}{2L^2}\right),$$

where $\lambda_{\max}(\boldsymbol{A})$ denotes the maximum eigenvalue of $\boldsymbol{A}$.

Note that the RHS is too optimistic: Suppose each $\boldsymbol{A}_i$ is diagonal with the diagonal containing independent entries. Then the diagonal of $\boldsymbol{A}$ contains d independent trials, and we can reasonably expect that in the worst case, applying (1.1) to each entry and taking a union bound will lose a factor of d.

As we will see, the conclusion is true with this correction:

$$\mathbb{P}\left[\lambda_{\max}(\boldsymbol{A}) \geqslant t\sqrt{n}\right] \leqslant d \cdot \exp\left(\frac{-t^2}{2L^2}\right). \tag{1.8}$$

Cautious optimism. So far everything is proceeding smoothly, so we might as well continue the analogy (cf. (1.2)):

$$\mathbb{P}\left[\lambda_{\max}(\boldsymbol{A}) \geqslant t\sqrt{n}\right] \leqslant \mathbb{P}\left[e^{\beta\lambda_{\max}(\boldsymbol{A})} \geqslant e^{\beta t\sqrt{n}}\right] \leqslant e^{-\beta t\sqrt{n}}\,\mathbb{E}\left[e^{\beta\lambda_{\max}(\boldsymbol{A})}\right].$$

Here we run into a hiccup: There is no nice formula for $\lambda_{\max}(\boldsymbol{A}_1 + \cdots + \boldsymbol{A}_n)$ in terms of $\lambda_{\max}(\boldsymbol{A}_1), \ldots, \lambda_{\max}(\boldsymbol{A}_n)$.

But since we willing to lose a factor of d in the tail bound (1.8), we might consider bounding the maximum eigenvalue by a sum:

$$\mathbb{P}\left[e^{\beta\lambda_{\max}(\boldsymbol{A})} \geqslant e^{\beta t\sqrt{n}}\right] \leqslant \mathbb{P}\left[\mathrm{Tr}(e^{\beta\boldsymbol{A}}) \geqslant e^{\beta t\sqrt{n}}\right], \tag{1.9}$$

where we recall the trace of a matrix:

$$\mathrm{Tr}(A) := \sum_{i=1}^{n} A_{ii}.$$

Let's take a moment to define e^X when X is a matrix.

Spectral functions. For a symmetric matrix A, let $\mathrm{spec}(A)$ be the set of A's eigenvalues. For any function $f : I \to \mathbb{R}$, we can extend f to symmetric matrices A with $\mathrm{spec}(A) \subseteq I$ as follows. Write $A = \sum_{i=1}^{d} \lambda_i u_i u_i^T$ where the $\{u_i\}$ are orthonormal, and define

$$f(A) := \sum_{i=1}^{d} f(\lambda_i) u_i u_i^T.$$

It is easily checked that $f(A)$ does not depend on the choice of an eigenvector basis $\{u_1, \ldots, u_d\}$ for A. When $f(t) = \sum_{j=1}^{m} c_j t^j$ is a polynomial, it holds that $f(A) = \sum_{j=1}^{m} c_j A^j$, as one might hope. This continues to hold when f has a convergent Taylor series. For $f(t) = e^t$, it holds that for all symmetric matrices A,

$$e^A := f(A) = \sum_{n\geqslant 0} \frac{A^n}{n!}.$$

For any $A, B \in \mathbb{M}_d$, a simple calculation gives

$$\mathrm{Tr}(AB) = \sum_{i,j} A_{ij} B_{ji} = \mathrm{Tr}(BA).$$

This is called the cyclic property of the trace. In particular, this implies that the trace is *invariant under similarilties:* For any invertible matrix P, we have

$$\mathrm{Tr}(PAP^{-1}) = \mathrm{Tr}(A).$$

For a symmetric matrix A with eigenvalues $\lambda_1, \ldots, \lambda_d$, we can diagonalize A and obtain $\mathrm{Tr}(A) = \lambda_1 + \cdots + \lambda_d$. This justifies (1.9).

Now we can apply Markov's inequality:

$$\mathbb{P}\left[\mathrm{Tr}\left(e^{\beta\boldsymbol{A}}\right) \geqslant e^{\beta t\sqrt{n}}\right] \leqslant e^{-\beta t\sqrt{n}}\,\mathbb{E}\left[\mathrm{Tr}\left(e^{\beta\boldsymbol{A}}\right)\right] = e^{-\beta t\sqrt{n}}\,\mathbb{E}\left[\mathrm{Tr}\left(e^{\beta(\boldsymbol{A}_1+\cdots+\boldsymbol{A}_n)}\right)\right]. \tag{1.10}$$

The trail goes cold. Unfortunately, non-commutativity now rears its head more seriously: In analogy with the classical setting, we want to utilize independence in evaluating a product. But we don't have a product: $e^{A+B} \neq e^A e^B$ unless A and B commute.

1.1 A sprinkle of magic

It's not surprising that we got stuck; many weaker, more difficult approaches were tried before the work of Ahlswede and Winter (2002). They employed the Golden-Thompson inequality.

Lemma 1.1 (Golden-Thompson)**.** *For all symmetric matrices $A, B \in \mathbb{M}_d$, it holds that*

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

And, indeed, it is not too difficult to establish our desired tail bound with this in hand.

$$\begin{aligned}\mathbb{E}\left[\operatorname{Tr}\left(e^{\beta(\boldsymbol{A}_1+\cdots+\boldsymbol{A}_n)}\right)\right] &\leqslant \mathbb{E}\left[\operatorname{Tr}\left(e^{\beta(\boldsymbol{A}_1+\cdots+\boldsymbol{A}_{n-1})}e^{\beta\boldsymbol{A}_n}\right)\right] \\ &= \mathbb{E}\left[\operatorname{Tr}\left(e^{\beta(\boldsymbol{A}_1+\cdots+\boldsymbol{A}_{n-1})}\,\mathbb{E}[e^{\beta\boldsymbol{A}_n}]\right)\right], \end{aligned} \tag{1.11}$$

where the last equality follows from independence (if $\boldsymbol{A}, \boldsymbol{B}$ are independent random matrices, it holds that $\mathbb{E}[\boldsymbol{AB}] = \mathbb{E}[\boldsymbol{A}]\,\mathbb{E}[\boldsymbol{B}]$).

For A, B both symmetric, it holds that

$$|\operatorname{Tr}(AB)| \leqslant \|A\| \cdot \|B\|_1. \tag{1.12}$$

where $\|A\| = \|(\lambda_1(A), \ldots, \lambda_d(A))\|_\infty$ denotes the operator norm, and $\|B\|_1 = \|(\lambda_1(B), \ldots, \lambda_d(B))\|_1$ is the ℓ_1-norm of the eigenvalues values of B. The proof of (1.12) is straightforward: Write $B = \sum_i \lambda_i u_i u_i^T$ where $\{u_i\}$ is an orthonormal basis. Then,

$$|\operatorname{Tr}(AB)| = \left|\sum_i \langle u_i, ABu_i\rangle\right| = \left|\sum_i \lambda_i \langle u_i, Au_i\rangle\right| \leqslant \|A\| \sum_i |\lambda_i| = \|A\| \cdot \|B\|_1.$$

When B is PSD, we have $\|B\|_1 = \operatorname{Tr}(B)$, hence (1.12) gives us

$$\mathbb{E}\left[\operatorname{Tr}\left(e^{\beta(\boldsymbol{A}_1+\cdots+\boldsymbol{A}_{n-1})}\,\mathbb{E}[e^{\beta\boldsymbol{A}_n}]\right)\right] \leqslant \left\|\mathbb{E}[e^{\beta\boldsymbol{A}_n}]\right\| \mathbb{E}\left[\operatorname{Tr}\left(e^{\beta(\boldsymbol{A}_1+\cdots+\boldsymbol{A}_{n-1})}\right)\right].$$

Continuing inductively, we end up with

$$\mathbb{E}\left[\operatorname{Tr}\left(e^{\beta(\boldsymbol{A}_1+\cdots+\boldsymbol{A}_n)}\right)\right] \leqslant \operatorname{Tr}(I)\prod_{i=1}^n \left\|\mathbb{E}[e^{\beta\boldsymbol{A}_i}]\right\| = d\prod_{i=1}^n \left\|\mathbb{E}[e^{\beta\boldsymbol{A}_i}]\right\| \tag{1.13}$$

Note that since $e^A = \sum_{n\geqslant 0} \frac{A^n}{n!}$, it holds that e^A commutes with powers of A. Since I, A, A^2, e^A all commute, they are simultaneously diagonalizable, meaning that the inequality (1.4) for real numbers extends:

$$I + A \preceq e^A \preceq I + A + A^2, \quad \forall \|A\| \leqslant 1.$$

Indeed, for diagonal matrices, this is just a pointwise inequality on the matrix entries: If D, D' are diagonal, then $D \preceq D' \iff D_{ii} \leqslant D'_{ii}\ \forall i$.

Therefore we have, for each i,

$$\left\|\mathbb{E}[e^{\beta\boldsymbol{A}_i}]\right\| \leqslant \|\,\mathbb{E}[1 + \beta\boldsymbol{A}_i + \beta^2\boldsymbol{A}_i]\| = \|\,\mathbb{E}[1 + \beta^2\boldsymbol{A}_i]\| \leqslant 1 + \beta^2 L^2 \leqslant e^{\beta^2 L^2}$$

where we have used the assumption $\|\boldsymbol{A}_i\| \leqslant L$. So combining (1.10) and (1.13) gives us

$$\mathbb{P}[\lambda_{\max}(\boldsymbol{A}) \geqslant t\sqrt{n}] \leqslant de^{-\beta t\sqrt{n}} e^{\beta^2 L^2 n},$$

which is analogous to the upper bound (1.5) we obtained in the scalar setting, except for the leading factor of d.