CSE 525: Randomized Algorithms **Spring 2023**

Lecture 12: Matrix Concentration Inequalities

Lecturer: Shayan Oveis Gharan *May 4th*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

12.1 Matrix Eigenvalues, A crash course

Let A be a $d \times d$ real symmetric matrix, then A has all real eigenvalues which we can order $\lambda_1(A) \geq \lambda_2(A) \geq \dots \geq \lambda_d(A)$. The operator norm of A is

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

The trace of A is

$$\text{Tr}(A) = \sum_{i=1}^{d} A_{i,i} = \sum_{i=1}^{n} \lambda_i(A).$$

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 $Tr(A) = \|A\|_*$. 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. Equivalently, if A has eigenvectors $v_1, \dots, v_d$ corresponding to $\lambda_1, \dots, \lambda_d$, then

$$e^A = \sum_{i=1}^{d} \lambda_i v_i v_i^T.$$

Observe that if A is symmetric, then e^A is always PSD, as the next argument shows. One often says A, e^A are simultaneously diagonalizable as they have the same set of eigenvectors.

Finally, note that for symmetric matrices A and B, we have $|\text{Tr}(AB)| \leq \|A\| \cdot \|B\|_*$. To see this, let $B = \sum_{i=1}^{d} \lambda_i u_i u_i^T$, then

$$|\text{Tr}(AB)| = \left|\sum_{i=1}^{d} \lambda_i \text{Tr}(A u_i u_i^T)\right| = \left|\sum_{i=1}^{d} \lambda_i u_i^T A u_i\right| \leq \left|\sum_{i=1}^{d} \lambda_i \|A\|\right| = \|A\| \|B\|_*.$$

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 12.1 (Golden-Thompson inequality)**.** *If A and B are real symmetric matrices, then* $\text{Tr}(e^{A+B}) \leq \text{Tr}(e^A e^B)$.

We do not prove this theorem here

12.2 The Laplace transform for matrices

We will consider now a random $d \times d$ real matrix X. The entries $X_{i,j}$ 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, recall that the trace is a linear operator and it commutes with expectation, namely $\mathrm{Tr}(\mathbb{E}[X]) = \mathbb{E}[\mathrm{Tr}(X)]$. Note that $\mathbb{E}[X]$ is the matrix defined by $(\mathbb{E}[X])_{i,j} = \mathbb{E}[X_{i,j}]$.

Suppose that $X_1, X_2, \ldots, X_n$ are independent random real symmetric matrices. Let $X = X_1 + X_2 + \cdots + X_n$. 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:

$$\mathbb{P}\left[\max_i \lambda_i(X) > t\right] = \mathbb{P}\left[\max_i e^{\beta \lambda_i(X)} > e^{\beta t}\right] \leq \mathbb{P}\left[\mathrm{Tr}(e^{\beta X}) > e^{\beta t}\right] \tag{12.1}$$

where the last inequality uses the fact that all the eigenvalues of $e^{\beta X}$ are nonnegative.

Now, Markov's inequality implies that

$$\mathbb{P}\left[\mathrm{Tr}(e^{\beta X} > e^{\beta t}\right] \leq \frac{\mathbb{E}\left[\mathrm{Tr}(e^{\beta X})\right]}{e^{\beta t}}. \tag{12.2}$$

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

Let $S_k = X_1 + \cdots + X_k$ be a prefix sum; so $S_n = X$. In the matrix setting, this is more subtle: Using Golden-Thmpson's inequality,

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

Now we push the expectation over X_n inside the trace:

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

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 $\mathrm{Tr}(AB) \leq \|A\| \cdot \|B\|_*$ and $\|B\|_* = \mathrm{Tr}(B)$ when B is PSD (as is the case for $e^{\beta S_{n-1}}$):

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

Now, by induction,

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

Combining with (12.1) and (12.2) we get

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

Applying the same thing to $-X$ we obtain,

$$\mathbb{P}\left[\|X\| > t\right] \leq e^{-\beta t} \cdot d \cdot \left(\prod_{i=1}^{n} \|\mathbb{E}\left[e^{\beta X_i}\right]\| + \prod_{i=1}^{n} \|\mathbb{E}\left[e^{-\beta X_i}\right]\|\right) \tag{12.3}$$

12.3 A Matrix Concentration Inequality

Let Y be a random, symmetric, psd $d \times d$ matrix with $\mathbb{E}[Y] = I$. Suppose that $\|Y\| \le L$ with probability one (this condition can be seen as analogue of Azuma-Hoeffding inequalities).

Theorem 12.2. *If $Y_1, Y_2, \ldots, Y_n$ are i.i.d. copies of Y, then for any $\epsilon \in (0,1)$ 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*

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

There is a slightly nicer way to write this using the Löwner ordering of symmetric matrices: Recall that $A \succeq B$ means that $A - B$ is PSD. We can rewrite the conclusion of the above theorem as

$$\mathbb{P}\left[(1-\epsilon)I \preceq \frac{1}{n}\sum_i Y_i \preceq (1+\epsilon)I\right] \le 1 - 2d\exp(-\epsilon^2 n/4L).$$

Proof. Define $X_i := Y_i - \mathbb{E}[Y_i]$ and $X = X_1 + \cdots + X_n$. Then the claim is equivalent to

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

We know from previous section that it will suffice to bound $\|\mathbb{E}\left[e^{\beta X_i}\right]\|$ for each i. First observe that, for all i,

$$\|X_i\| = \|Y_i - \mathbb{E}[Y_i]\| \underset{Y_i \succeq 0}{=} \|Y_i\| - 1 = L - 1.$$

So, for $\beta < 1/L$, we have $-I \preceq \beta X_i \preceq I$. Next, we use the fact that

$$1 + x \le e^x \le 1 + x + x2 \forall x \in [-1, 1].$$

Note that if A is a real symmetric matrix with $\|A\| \le 1$, then since I, A, A^2, and e^A are simultaneously diagonalizable, this yields

$$I + A \preceq e^A \preceq I + A + A^2.$$

So,

$$\mathbb{E}\left[e^{\beta X_i}\right] \preceq I + \beta\mathbb{E}[X_i] + \beta^2\mathbb{E}\left[X_i^2\right] \underset{\mathbb{E}[X_i]=0}{=} I + \beta^2\mathbb{E}\left[X_i^2\right] \preceq e^{\beta^2\mathbb{E}\left[X_i^2\right]}$$

Lastly,

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

Therefore, $\|\mathbb{E}\left[e^{\beta X_i}\right]\| \le \|e^{\beta^2 LI}\| = e^{\beta^2 L}$. Plugging this back into (12.3) we get

$$\mathbb{P}\left[\|X\| > t\right] \le 2de^{-\beta t}e^{\beta^2 Ln} \underset{\beta=\epsilon/L}{\le} 2de^{-\epsilon^2 n/4L}.$$

This completes the proof. □