Matrix analogues of certain large-deviation inequalities

Sparsho De

May 24, 2024

1 Introduction

As shown in class, a particularly critical idea for proving large deviation inequalities is Chernoff bounds.

Theorem 1.1 (Chernoff's Inequality). *Let* $X_1, X_2, X_3, \ldots, X_n \sim X$ *where* $\mathbb{E}[X] = 0$ *and* $|X| < 1$ *almost surely. Then we have*

$$\mathbb{P}(X_1 + X_2 + \cdots + X_n \geq \lambda\sigma) \leq \max(e^{-\lambda^2/4}, e^{-\lambda\sigma/2}).$$

Proof. Letting $S_n = X_1 + \cdots + X_n$, we have

$$p = \mathbb{P}(S_n > t) = \mathbb{P}(e^{\lambda S_n} > e^{\lambda t} \leq e^{-\lambda t} \prod_i \mathbb{E}e^{\lambda X_i}$$

by independence of X_i and Markov's Inequality. Furthermore, since $\mathbb{E}[X_i] = 0$ and $|X_i| < 1$ by Taylor-series expansion, we have

$$\mathbb{E}e^{\lambda X_i} \lesssim e^{\lambda^2 \operatorname{Var} X_i}.$$

This yields

$$p \lesssim e^{-\lambda t + \lambda^2 \sigma^2}$$

with $\sigma^2 = \sum_{i=1}^n \operatorname{Var} X_i$. It suffices to check optimize over the parameter λ. After differentiating and setting equal to zero, we find that the optimal $\lambda = \min(t^2/2\sigma^2, 1)$ which gives us our main result. $\square$

This result underpins many of the other concentration inequalities proved in class. On the other hand, it does not easily extend to proving matrix inequalities. In particular, consider two matrices A, B. If $A < B$, that is, $B - A$ is positive semi-definite, then it is not immediately clear that $e^A < e^B$. Then, we cannot simply replicate the proof of Chernoff's inequality to get a matrix analogue. There is some more work to be done.

2 Golden-Thompson Inequality

Therefore, the critical step to transform large-deviation inequalities from scalar random variables to *non-commutative* matrix random variables is the Golden-Thompson Inequality. [1]

Theorem 2.1 (Golden-Thompson Inequality)**.**

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

Indeed, this result is somewhat surprising. If $AB = BA$ we know that $e^{A+B} = e^A e^B$. However, it is not immediately clear if there is a relationship between e^A and e^B for non-commutative matrices. For the proof, we follow Dyson's work in 1965.

First, we establish a few lemmas.

Lemma 2.2. *For any square matrices X, Y we have that*

$$|\,\mathrm{Tr}(XY)|^2 \leq \mathrm{Tr}(X^T X)\,\mathrm{Tr}(Y^T Y).$$

Proof. This is just Cauchy-Schwartz inequality on the trace inner product. □

Lemma 2.3. *Let P be any product of $2n$ factors which may be X or X^T in any order. Then,*

$$|\,\mathrm{Tr}(P)| \leq \mathrm{Tr}(XX^T)^n.$$

Proof. Among all the choices of P, pick the one that maximizes $|\,\mathrm{Tr}(P)|$. Obviously, if P is of the form $(X^T X)^n$ or $(XX^T)^n$ we are done. So, suppose it is not of that form. Then, there exists a pair of consecutive factors of X and X^T. Since $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ we can simply permute the entries of P such that there is a X and X^T on the n-th and $(n+1)$-th index respectively. Now, write $P = QR$ with Q being the product of the first n terms and Q being the product of the last n terms. Apply 2.2 we have

$$|\,\mathrm{Tr}(P)|^2 \leq \mathrm{Tr}(Q^T Q)\,\mathrm{Tr}(R^T R).$$

However, since $P' = Q^T Q$ and $P'' = P^T P$ are of the same form as P we obviously have $|\,\mathrm{Tr}(P')| \leq |\,\mathrm{Tr}(P)|$ and also $|\,\mathrm{Tr}(P'')| \leq |\,\mathrm{Tr}(P)|$. Therefore, we have the equality

$$|\,\mathrm{Tr}(P)| = |\,\mathrm{Tr}(P')| = |\,\mathrm{Tr}(P'')|.$$

Let k, k', and k'' denote the number of neighbor XX^T pairs in P, P', and P'' respectively. In particular, we count the last and first entry as neighbors. Note, we have two cases.

$$\begin{cases} k' + k'' = 2k + 1 & \text{if first and last factors of } P \text{ are different} \\ k' + k'' = 2k + 2 & \text{if first and last factors of } P \text{ are the same} \end{cases}.$$

By Pigeonhole principle, we must have that *at least one* of either P' or P'' as more XX^T neighbor-pairs than P. Now, consider the P that attains the maximum $|\operatorname{Tr}(P)|$ and maximizes the number of XX^T pairs. Applying the argument, the only case where at least one of P' and P'' has more neighbor pairs than P is when $P = (X^T X)^n$ or when $P = (XX^T)^n$. □

Lemma 2.4. *For any two Hermitian matrices A and B we have that*

$$\operatorname{Tr}(A^{2^k} B^{2^k}) \geq \operatorname{Tr}(AB)^{2^k}.$$

Proof. We just apply 2.3. Taking $X = AB$ and $X^T = BA$ we have

$$|\operatorname{Tr}(AB)^{2n}| \leq \operatorname{Tr}(ABBA)^n = \operatorname{Tr}(A^2B^2)^n.$$

Now, take $X = A^2B$. So, we have

$$|\operatorname{Tr}(AB)^{4n}| \leq \operatorname{Tr}(A^2B^2B^2A^2)^n = \operatorname{Tr}(A^4B^4)^n.$$

We can just inductively repeat this argument to get our result. □

Using the previous lemma, we can immediately prove the Golden-Thompson inequality.

Proof of Theorem 2.1. Just take $A' = (1 + 2^{-k}A)$ and $B' = (1 + 2^{-k}B)$ and apply 2.4. Taking the limit as $k \to \infty$ we have our result. □

This resolves the main issue with transforming the proof of Chernoff's inequality for scalars to a Chernoff's inequality for matrices. Hence, we are ready to prove the matrix analog. The method of proof is similar to the original.

3 Matrix analog for Chernoff's Inequality

Before we can begin a proof, we have remind the reader of some notation. In particular, we have the partial order for square matrices A, B with $A \leq B$ implying that $A - B$ is positive semi-definite.

Theorem 3.1 (Chernoff-type inequality)**.** *Let M_d denote the class of symmetric $d \times d$ matrices. Let $X_i \in M_d$ be independent mean zero random matrices, $||X_i|| \leq 1$ for all i almost surely. Let $S_n = X_1 + \dots X_n$ and $\sigma^2 = \sum_{n=1}^{\infty} ||\operatorname{Var}(X_i)||$. Then, for every $t > 0$ we have*

$$\mathbb{P}(||S_n|| > t) \leq d \cdot \max(e^{-t^2/4\sigma^2}, e^{-t/2}).$$

However, before we can prove the theorem, we need a way to estimate $\mathbb{E}e^Z$ for an arbitrary mean zero random matrix Z with $||Z|| < 1$. The reason is identical to the reason for the original proof of Chernoff's inequality. Therefore, we have the following lemma.

Lemma 3.2. *Let $Z \in M_d$ be a mean zero random matrix, $||Z|| < 1$ a.s. Then,*

$$\mathbb{E}e^Z \leq e^{\operatorname{Var} Z}.$$

Proof. Just like the original proof of this statement for real random variables, we apply Taylor-series expansion. Indeed, we have

$$\mathbb{E}e^Z \leq \mathbb{E}(I + Z + Z^2) = I + \mathbb{E}(Z) + \mathbb{E}(Z^2) = I + \mathrm{Var}(Z) \leq e^{\mathrm{Var}(Z)}.$$

$\square$

Now, we are ready to prove the matrix analog of Chernoff's inequality.

Proof of Theorem 3.1. Note,

$$\begin{aligned} p = \mathbb{P}(S_n \not\leq tI) = \mathbb{P}(e^{\lambda S_n} \not\leq e^{\lambda t I}) &\leq \mathbb{P}(\mathrm{Tr}(e^{\lambda S_n}) > e^{\lambda t}) \\ &\leq e^{-\lambda t}\mathbb{E}\,\mathrm{Tr}(e^{\lambda S_n}) \end{aligned}$$

with the last step following from Markov's inequality. So now, it suffices to estimate $\mathbb{E}\,\mathrm{Tr}(e^{\lambda S_n})$. Since $S_n = X_n + S_{n-1}$ we use 2.1 to see that

$$\mathbb{E}\,\mathrm{Tr}(e^{S_n}) \leq \mathbb{E}\,\mathrm{Tr}(e^{\lambda X_n} e^{\lambda S_{n-1}}).$$

Now, using that X_n and S_{n-1} are independent, along with the fact that $\mathbb{E}$ and Tr commute, we have that this is equal to

$$\mathbb{E}(\mathrm{Tr}(\mathbb{E}(e^{\lambda X_n} e^{\lambda S_{n-1}}))) \leq ||\mathbb{E}e^{\lambda X_n}|| \cdot \mathbb{E}\,\mathrm{Tr}(e^{\lambda S_{n-1}}).$$

We can inductively repeat this process. Using the fact that $\mathrm{Tr}(I) = \mathrm{Tr}(I_d) = d$ we have

$$\mathbb{E}\,\mathrm{Tr}(e^{\lambda S_n}) \leq d \prod_{i=1}^{n} ||\mathbb{E}e^{\lambda X_i}||.$$

Therefore, we have shown

$$\mathbb{P}(S_n \not\leq tI) \leq de^{-\lambda t} \prod_{i=1}^{n} ||\mathbb{E}e^{\lambda X_i}||.$$

Simply repeating the process for $-S_n$ and using the fact that $tI_d \leq S_n \leq tI_d \iff ||S_n|| \leq t$ we have our main result that

$$\mathbb{P}(||S_n|| > t) \leq 2d^{-\lambda t} \cdot \prod_{i=1}^{n} ||\mathbb{E}e^{\lambda X_i}||.$$

But, using 3.2 we can easily estimate this quantity. Indeed, we have $||\mathbb{E}e^{\lambda X_i}|| \leq ||e^{\lambda^2 \mathrm{Var}(X_i)}|| = e^{\lambda^2 || \mathrm{Var}(X_i)||}$. Hence, we have the result

$$\mathbb{P}(||S|| > t) \leq d \cdot e^{-\lambda t + \lambda^2 \sigma^2}.$$

It suffices to optimize over λ. Simply differentiating with respect to λ and optimizing, we have $\lambda = \min(t/2\sigma^2, 1)$. This gives us our main result. $\square$

Indeed, we have an immediate corollary.

Theorem 3.3. *Let $X_i \in M_d$ be independent random matrices $X_i \geq 0, ||X_i|| \leq 1$ for all i almost surely. As usual, let $S_n = X_1 + \cdots + X_n$ and $E = \sum_{i=1}^n ||\mathbb{E}X_i||$. Then for every $\epsilon \in (0,1)$ we have*

$$\mathbb{P}(||S_n - \mathbb{E}S_n|| > \epsilon E) \leq d \cdot e^{-\epsilon^2 E/4}.$$

Proof. This is basically an immediate application of 3.1. That is, applying the theorem for $X_i - \mathbb{E}X_i$ we have

$$\mathbb{P}(||S_n - \mathbb{E}S_n|| > \epsilon E) \leq d \cdot \max(e^{-t^2/4\sigma^2}, e^{-t/2}).$$

It suffices to bound this right-hand term. Note that $||X_i|| \leq 1 \implies \mathrm{Var}(X_i) \leq \mathbb{E}X_i^2 \leq \mathbb{E}(||X_i||X_i) \leq \mathbb{E}(X_i)$. Therefore, we have that $\sigma^2 \leq E$. Now, just replace $t = \epsilon E$. We have that

$$t^2/4\sigma^2 = \epsilon^2 E^2/4\sigma^2 \geq \epsilon^2 E/4.$$

Hence, we have our main result. □

4 Matrix analog for Khintchine's inequality

For this, we follow Oliveira's work [2]. We share a **short** outline of the proof rather than the details, since much of the work is similar to above.

Theorem 4.1 (Khintchine-type inequality)**.** *Given positive integers $d, n \in \mathbb{N}$ let $A_1, \ldots, A_n$ be deterministic $d \times d$ Hermitian matrices and $\{\epsilon_i\}_{i=1}^n$ be an i.i.d sequence of Rademacher random variables. Define $Z_n = \sum_{i=1}^n \epsilon_i A_i$. Then, for all $p \in [1, +\infty)$, we have*

$$\mathbb{E}[||Z_n||^p]^{1/p} \leq (\sqrt{2\log(2d)} + c\sqrt{p})||\sum_{i=1}^n A_i^2||^{1/2}$$

for a independent constant c.

Proof outline. We wish to control the tail behavior of $||Z_n||$. Using a similar idea to above, we can show

$$\mathbb{P}(||Z_n|| > t) \leq 2 \inf_{\lambda > 0} e^{-\lambda t} \mathbb{E}\,\mathrm{Tr}(e^{\lambda Z_n}).$$

It suffices to control this right term. Just like before, using 2.1 we have

$$\mathbb{E}\,\mathrm{Tr}(e^{\lambda Z_n}) \leq \mathrm{Tr}\left(e^{\frac{\lambda^2 \sum_{i=1}^n A_i^2}{2}}\right).$$

Combining these two results together, we have

$$\mathbb{P}(||Z_n|| > t) \leq 2de^{-t^2/2\sigma^2}.$$

Simply considering the equation

$$\frac{1}{\sigma^p}\mathbb{E}\left[(||Z_n|| - \sqrt{2\log(2d}\sigma)^p\right]$$

and doing some calculus gives us our result.

□

References

[1] Peter J. Forrester and Colin J. Thompson. The golden-thompson inequality: Historical aspects and random matrix applications. *Journal of Mathematical Physics*, 55(2), February 2014.

[2] Roberto Imbuzeiro Oliveira. Sums of random hermitian matrices and an inequality by rudelson, 2010.