Sums of random Hermitian matrices and an inequality by Rudelson

Roberto Imbuzeiro Oliveira*

April 23, 2010

Abstract

We give a new, elementary proof of a key inequality used by Rudelson in the derivation of his well-known bound for random sums of rank-one operators. Our approach is based on Ahlswede and Winter's technique for proving operator Chernoff bounds. We also prove a concentration inequality for sums of random matrices of rank one with explicit constants.

1 Introduction

This note mainly deals with estimates for the operator norm $\|Z_n\|$ of random sums

$$Z_n \equiv \sum_{i=1}^{n} \epsilon_i A_i \tag{1}$$

of deterministic Hermitian matrices $A_1, \dots, A_n$ multiplied by random coefficients. Recall that a *Rademacher sequence* is a sequence $\{\epsilon_i\}_{i=1}^n$ of i.i.d. random variables with ϵ_1 uniform over $\{-1, +1\}$. A *standard Gaussian sequence* is a sequence i.i.d. standard Gaussian random variables. Our main goal is to prove the following result.

Theorem 1 (proven in Section 3) *Given positive integers $d, n \in \mathbb{N}$, let $A_1, \dots, A_n$ be deterministic $d \times d$ Hermitian matrices and $\{\epsilon_i\}_{i=1}^n$ be either a Rademacher sequence or a standard Gaussian sequence. Define Z_n as in (1). Then for all $p \in [1, +\infty)$,*

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

*IMPA, Rio de Janeiro, RJ, Brazil, 22430-040. rimfo@impa.br

where

$$C_p \equiv \left(p\int_0^{+\infty} t^{p-1}e^{-\frac{t^2}{2}}\,dt\right)^{1/p} \quad (\leq c\sqrt{p} \textit{ for some universal } c > 0).$$

For $d = 1$, this result corresponds to the classical Khintchine inequalities, which give sub-Guassian bounds for the moments of $\sum_{i=1}^n \epsilon_i a_i$ $(a_1, \dots, a_n \in \mathbb{R})$. Theorem 1 is implicit in Section 3 of Rudelson's paper [11], albeit with non-explicit constants. The main Theorem in that paper is the following inequality, which is a simple corollary of Theorem 1: if $Y_1, \dots, Y_n$ are i.i.d. random (column) vectors in $\mathbb{C}^d$ which are isotropic (i.e $\mathbb{E}[Y_1 Y_1^*] = I$, the $d \times d$ identity matrix), then:

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n Y_i Y_i^* - I\right\|\right] \leq C\,\mathbb{E}\left[|Y_1|^{\log n}\right]^{1/\log n}\sqrt{\frac{\log d}{n}} \tag{2}$$

for some universal $C > 0$, whenever the RHS of the above inequality is at most 1. This important result has been applied to several different problems, such as bringing a convex body to near-isotropic position [11]; the analysis of for low-rank approximations of matrices [12, 6] and graph sparsification [13]; estimating of singular values of matrices with independent rows [10]; analysing compressive sensing [3]; and related problems in Harmonic Analysis [16, 15].

The key ingredient of the original proof of Theorem 1 is a non-commutative Khintchine inequality by Lust-Picard and Pisier [9]. This states that there exists a universal $c > 0$ such that for all Z_n as in the Theorem, all $p \geq 1$ and all $d \times d$ matrices $\{B_i, D_i\}_{i=1}^n$ with $B_i + D_i = A_i$, $1 \leq i \leq n$,

$$\mathbb{E}\left[\|Z_n\|_{S^p}^p\right]^{1/p} \leq c\sqrt{p}\left(\left\|\sum_{i=1}^n B_i B_i^*\right\|_{S^p}^{1/2} + \left\|\sum_{i=1}^n D_i^* D_i\right\|_{S^p}^{1/2}\right),$$

where $\|\cdot\|_{S^p}$ denotes the *p-th Schatten norm*: $\|A\|_{S^p}^p \equiv \mathrm{Tr}[(A^*A)^{p/2}]$. Unfortunately, the proof of the Lust-Picard/Pisier inequality employs language and tools from non-commutative probability that are rather foreign to most potential users of (2).

This note presents an elementary proof of Theorem 1 that bypasses the above inequality. Our argument is based on an improvement of the methodology created by Ahlswede and Winter [2] in order to prove their *operator Chernoff bound*, which also has many applications e.g. [7] (the improvement is discussed in Section 3.1). This approach only requires elementary facts from Linear Algebra and Matrix Analysis. The most complicated result that we use is the Golden-Thompspon inequality [5, 14]:

$$\forall d \in \mathbb{N},\ \forall\ d \times d \text{ Hermitian matrices } A, B,\ \mathrm{Tr}(e^{A+B}) \leq \mathrm{Tr}(e^A e^B). \tag{3}$$

The elementary proof of this classical inequality is sketched in Section 5 below.

We have already noted that Rudelson's bound (2) follows simply from Theorem 1; see [11, Section 3] for detais. Here we prove a concentration lemma corresponding to that result under the stronger assumption that $|Y_1|$ is a.s. bounded. While similar results have appeared in other papers [10, 12, 16], our proof is simpler and gives explicit (albeit quite large) constants.

Lemma 1 (Proven in Section 4) *Let $Y_1, \dots, Y_n$ be i.i.d. random column vectors in $\mathbb{C}^d$ with $|Y_1| \leq M$ almost surely and $\|\mathbb{E}[Y_1 Y_1^*]\| \leq 1$. Then:*

$$\forall t \geq 0, \mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n} Y_i Y_i^* - \mathbb{E}[Y_1 Y_1^*]\right\| \geq t\right) \leq (2n)^2 e^{-\frac{nt^2}{16M^2+8M^2t}}.$$

In particular, a calculation shows that:

$$\left\|\frac{1}{n}\sum_{i=1}^{n} Y_i Y_i^* - \mathbb{E}[Y_1 Y_1^*]\right\| < \epsilon(n, M) \equiv M\sqrt{\frac{72\ln n + 48\ln 2}{n}} \text{ with probability } \geq 1 - \frac{1}{n}$$

whenever $\epsilon(n, M) \leq 1$. A key feature both of this Lemma is that the ambient dimension d plays no direct role in the bound. In fact, the same result holds for Y_i taking values in a separable Hilbert space (as in the last section of [10]).

To conclude the introduction, we present an open problem: is it possible to improve upon Rudelson's bound under further assumptions? There is some evidence that the dependence on $\ln(d)$ in the Theorem, while necessary in general [12, Remark 3.4], can sometimes be removed. For instance, Adamczak et al. [1] have improved upon Rudelson's original application of Theorem 1 to convex bodies, obtaining exactly what one would expect in the absence of the $\sqrt{\log(2d)}$ term. Another setting where our bound is a $\Theta\left(\sqrt{\ln d}\right)$ factor away from optimality is that of more classical random matrices (cf. the end of Section 3.1 below). It would be interesting if one could sharpen the proof of Theorem 1 in order to reobtain these results. [Related issues are raised by Vershynin [17].]

2 Preliminaries

We let $\mathbb{C}^{d\times d}_{\rm Herm}$ denote the set of $d \times d$ Hermitian matrices, which is a subset of the set $\mathbb{C}^{d\times d}$ of all $d \times d$ matrices with complex entries. The *spectral theorem* states that all $A \in \mathbb{C}^{d\times d}_{\rm Herm}$ have d real eigenvalues (possibly with repetitions) that correspond to an orthonormal set of eigenvectors. $\lambda_{\max}(A)$ is the largest eigenvalue of A. The spectrum of A, denoted by spec(A), is the multiset of all eigenvalues, where each eigenvalue appears a number of times equal to its multiplicity. We let

$$\|C\| \equiv \max_{v\in\mathbb{C}^d\,|v|=1} |Cv|$$

denote the operator norm of $C \in \mathbb{C}^{d\times d}$ ($|\cdot|$ is the Euclidean norm). By the spectral theorem,

$$\forall A \in \mathbb{C}^{d\times d}_{\rm Herm},\ \|A\| = \max\{\lambda_{\max}(A), \lambda_{\max}(-A)\}.$$

Moreover, $\mathrm{Tr}(A)$ (the trace of A) is the sum of the eigenvalues of A.

2.1 Spectral mapping

Let $f : \mathbb{C} \to \mathbb{C}$ be an entire analytic function with a power-series representation $f(z) \equiv \sum_{n\geq 0} c_n\, z^n$ ($z \in \mathbb{C}$). If all c_n are real, the expression:

$$f(A) \equiv \sum_{n\geq 0} c_n A^n \ \ (A \in \mathbb{C}^{d\times d}_{\rm Herm})$$

corresponds to a map from $\mathbb{C}^{d\times d}_{\rm Herm}$ to itself. We will sometimes use the so-called spectral mapping property:

$$\mathrm{spec} f(A) = f(\mathrm{spec}(A)). \tag{4}$$

By this we mean that the eigenvalues of $f(A)$ are the numbers $f(\lambda)$ with $\lambda \in \mathrm{spec}(A)$. Moreover, the multiplicity of $\xi \in \mathrm{spec} f(A)$ is the sum of the multiplicities of all preimages of ξ under f that lie in spec(A).

2.2 The positive-semidefinite order

We will use the notation $A \succeq 0$ to say that A is *positive-semidefinite*, i.e. $A \in \mathbb{C}^{d\times d}_{\rm Herm}$ and its eigenvalues are A are non-negative. This is equivalent to saying that $(v, Av) \geq 0$ for all $v \in \mathbb{C}^d$, where $(\cdot, \cdot\cdot)$ is the standard Euclidean inner product.

If $A, B \in \mathbb{C}^{d\times d}_{\rm Herm}$, we write $A \succeq B$ or $B \preceq A$ to say that $A - B \succeq 0$. Notice that "$\succeq$" is a partial order and that:

$$\forall A, B, A', B' \in \mathbb{C}^{d\times d}_{\rm Herm},\ (A \preceq A') \wedge (B \preceq B') \Rightarrow A + A' \preceq B + B'. \tag{5}$$

Moreover, spectral mapping (4) implies that:

$$\forall A \in \mathbb{C}^{d\times d}_{\rm Herm},\ A^2 \succeq 0. \tag{6}$$

We will also need the following simple fact.

Proposition 1 *For all* $A, B, C \in \mathbb{C}^{d\times d}_{\rm Herm}$:

$$(C \succeq 0) \wedge (A \preceq B) \Rightarrow \mathrm{Tr}(AC) \leq \mathrm{Tr}(BC). \tag{7}$$

Proof: To prove this, assume the LHS and observe that the RHS is equivalent to $\mathrm{Tr}(C\Delta) \geq 0$ where $\Delta \equiv B - A$. By assumption, $\Delta \succeq 0$, hence it has a Hermitian square root $\Delta^{1/2}$. The cyclic property of the trace implies:

$$\mathrm{Tr}(C\Delta) = \mathrm{Tr}(\Delta^{1/2}C\Delta^{1/2}).$$

Since the trace is the sum of the eigenvalues, we will be done once we show that $\Delta^{1/2}C\Delta^{1/2} \succeq 0$. But, since $\Delta^{1/2}$ is Hermitian and $C \succeq 0$,

$$\forall v \in \mathbb{C}^d,\ (v, \Delta^{1/2}C\Delta^{1/2}v) = ((\Delta^{1/2}v), C(\Delta^{1/2}v)) = (w, Cw) \geq 0 \text{ (with } w = \Delta^{1/2}v),$$

which shows that $\Delta^{1/2}C\Delta^{1/2} \succeq 0$, as desired. □

2.3 Probability with matrices

Assume $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $Z : \Omega \to \mathbb{C}^{d\times d}_{\rm Herm}$ is measurable with respect to $\mathcal{F}$ and the Borel σ-field on $\mathbb{C}^{d\times d}_{\rm Herm}$ (this is equivalent to requiring that all entries of Z be complex-valued random variables). $\mathbb{C}^{d\times d}_{\rm Herm}$ is a metrically complete vector space and one can naturally define an expected value $\mathbb{E}[Z] \in \mathbb{C}^{d\times d}_{\rm Herm}$. This turns out to be the matrix $\mathbb{E}[Z] \in \mathbb{C}^{d\times d}_{\rm Herm}$ whose (i, j)-entry is the expected value of the (i, j)-th entry of Z. [Of course, $\mathbb{E}[Z]$ is only defined if all entries of Z are integrable, but this will always be the case in this paper.]

The definition of expectations implies that traces and expectations commute:

$$\mathrm{Tr}(\mathbb{E}[Z]) = \mathbb{E}[\mathrm{Tr}(Z)]. \tag{8}$$

Moreover, one can check that the usual product rule is satisfied:

$$\text{If } Z, W : \Omega \to \mathbb{C}^{d\times d}_{\rm Herm} \text{ are measurable and independent, } \mathbb{E}[ZW] = \mathbb{E}[Z]\,\mathbb{E}[W]. \tag{9}$$

3 Proof of Theorem 1

Proof: [of Theorem 1] We wish to control the tail behavior of:

$$\|Z_n\| = \max\{\lambda_{\max}(Z_n), \lambda_{\max}(-Z_n)\}.$$

However, Z_n and $-Z_n$ have the same distribution. It follows that:

$$\forall t \geq 0,\, \mathbb{P}\left(\|Z_n\| \geq t\right) \leq 2\,\mathbb{P}\left(\lambda_{\max}(Z_n) \geq t\right).$$

The usual Bernstein trick implies that for all $t \geq 0$,

$$\forall t \geq 0,\, \mathbb{P}\left(\lambda_{\max}(Z_n) \geq t\right) \leq \inf_{s>0} e^{-st}\mathbb{E}\left[e^{s\lambda_{\max}(Z_n)}\right].$$

The function "$x \mapsto e^{sx}$" is monotone non-decreasing and positive for all $s \geq 0$. It follows from the spectral mapping property (4) that for all $s \geq 0$, the largest eigenvalue of e^{sZ_n} is $e^{s\lambda_{\max}(Z_n)}$ and all eigenvalues of e^{sZ_n} are non-negative. Using the equality "trace = sum of eigenvalues" implies that for all $s \geq 0$,

$$\mathbb{E}\left[e^{s\lambda_{\max}(Z_n)}\right] = \mathbb{E}\left[\lambda_{\max}\left(e^{sZ_n}\right)\right] \leq \mathbb{E}\left[\mathrm{Tr}\left(e^{sZ_n}\right)\right].$$

As a result, we have the inequality:

$$\forall t \geq 0,\, \mathbb{P}\left(\|Z_n\| \geq t\right) \leq 2\inf_{s\geq 0} e^{-st}\mathbb{E}\left[\mathrm{Tr}\left(e^{sZ_n}\right)\right]. \tag{10}$$

Up to now, our proof has followed Ahlswede and Winter's argument. The next lemma, however, will require new ideas.

Lemma 2 *For all $s \in \mathbb{R}$,*

$$\mathbb{E}\left[\mathrm{Tr}(e^{sZ_n})\right] \leq \mathrm{Tr}\left(e^{\frac{s^2\sum_{i=1}^n A_i^2}{2}}\right).$$

This lemma is proven below. We will now show how it implies Rudelson's bound. Let

$$\sigma^2 \equiv \left\|\sum_{i=1}^n A_i^2\right\| = \lambda_{\max}\left(\sum_{i=1}^n A_i^2\right).$$

[The second inequality follows from $\sum_{i=1}^n A_i^2 \succeq 0$, which holds because of (5) and (6).] We note that:

$$\mathrm{Tr}\left(e^{\frac{s^2\sum_{i=1}^n A_i^2}{2}}\right) \leq d\,\lambda_{\max}\left(e^{\frac{s^2\sum_{i=1}^n A_i^2}{2}}\right) = d\,e^{\frac{s^2\sigma^2}{2}}$$

where the equality is yet another application of spectral mapping (4) and the fact that "$x \mapsto e^{s^2x/2}$" is monotone increasing. We deduce from the Lemma and (10) that:

$$\forall t \geq 0,\, \mathbb{P}\left(\|Z_n\| \geq t\right) \leq 2d\inf_{s\geq 0} e^{-st+\frac{s^2t^2}{2}} = 2d\,e^{-\frac{t^2}{2\sigma^2}}. \tag{11}$$

This implies that for any $p \geq 1$,

$$\begin{aligned}
\frac{1}{\sigma^p}\mathbb{E}\left[(\|Z_n\| - \sqrt{2\ln(2d)}\sigma)_+^p\right] &= p\int_0^{+\infty} t^{p-1}\mathbb{P}\left(\|Z_n\| \geq (\sqrt{2\ln(2d)}+t)\sigma\right)\, dt \\
(\text{use}(11)) \quad &\leq 2pd\int_0^{+\infty} t^{p-1}e^{-\frac{(t+\sqrt{2\ln(2d)})^2}{2}}\, dt \\
&\leq 2pd\int_0^{+\infty} t^{p-1}e^{-\frac{t^2+2\ln(2d)}{2}}\, dt \; = C_p^p
\end{aligned}$$

Since $0 \leq \|Z_n\| \leq \sqrt{2\ln(2d)}\sigma + (\|Z_n\| - \sqrt{2\ln(2d)}\sigma)_+$, this implies the L^p estimate in the Theorem. The bound "$C_p \leq c\sqrt{p}$" is standard and we omit its proof. $\square$

To finish, we now prove Lemma 2.

Proof: [of Lemma 2] Define $D_0 \equiv \sum_{i=1}^n s^2A_i^2/2$ and

$$D_j \equiv D_0 + \sum_{i=1}^{j}\left(s\epsilon_i A_i - \frac{s^2A_i^2}{2}\right) \quad (1 \leq j \leq n).$$

We will prove that for all $1 \leq j \leq n$:

$$\mathbb{E}\left[\mathrm{Tr}\left(\exp\left(D_j\right)\right)\right] \leq \mathbb{E}\left[\mathrm{Tr}\left(\exp\left(D_{j-1}\right)\right)\right]. \tag{12}$$

Notice that this implies $\mathbb{E}\left[\mathrm{Tr}(e^{D_n})\right] \leq \mathbb{E}\left[\mathrm{Tr}(e^{D_0})\right]$, which is the precisely the Lemma. To prove (12), fix $1 \leq j \leq n$. Notice that D_{j-1} is independent from $s\epsilon_jA_j - s^2A_j^2/2$ since the $\{\epsilon_i\}_{i=1}^n$ are independent. This implies that:

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Tr}\left(\exp\left(D_j\right)\right)\right] &= \mathbb{E}\left[\mathrm{Tr}\left(\exp\left(D_{j-1} + s\epsilon_jA_j - \frac{s^2A_j^2}{2}\right)\right)\right] \\
(\text{use Golden-Thompson (3)}) \quad &\leq \mathbb{E}\left[\mathrm{Tr}\left(\exp\left(D_{j-1}\right)\exp\left(s\epsilon_jA_j - \frac{s^2A_j^2}{2}\right)\right)\right] \\
(\mathrm{Tr}(\cdot) \text{ and } \mathbb{E}\left[\cdot\right] \text{ commute, (8)}) \quad &= \mathrm{Tr}\left(\mathbb{E}\left[\exp\left(D_{j-1}\right)\exp\left(s\epsilon_jA_j - \frac{s^2A_j^2}{2}\right)\right]\right). \\
(\text{use product rule, (9)}) \quad &= \mathrm{Tr}\left(\mathbb{E}\left[\exp\left(D_{j-1}\right)\right]\mathbb{E}\left[\exp\left(s\epsilon_jA_j - \frac{s^2A_j^2}{2}\right)\right]\right).
\end{aligned}$$

By the monotonicity of the trace (7) and the fact that $\exp\left(D_{j-1}\right) \succeq 0$ (which follows from (4)), we will be done once we show that:

$$\mathbb{E}\left[\exp\left(s\epsilon_jA_j - \frac{s^2A_j^2}{2}\right)\right] \preceq I. \tag{13}$$

The key fact is that $s\epsilon_j A_j$ and $-s^2A_j^2/2$ *always* commute, hence the exponential of the sum is the product of the exponentials. Applying (9) and noting that $e^{-s^2A_j^2/2}$ is constant, we see that:

$$\mathbb{E}\left[\exp\left(s\epsilon_j A_j - \frac{s^2A_j^2}{2}\right)\right] = \mathbb{E}\left[\exp\left(s\epsilon_j A_j\right)\right] e^{-\frac{s^2A_j^2}{2}}.$$

In the Gaussian case, an explicit calculation shows that $\mathbb{E}\left[\exp\left(s\epsilon_j A_j\right)\right] = e^{s^2A_j^2/2}$, hence (13) holds. In the Rademacher case, we have:

$$\mathbb{E}\left[\exp\left(s\epsilon_j A_j\right)\right] e^{-\frac{s^2A_j^2}{2}} = f(A_j)$$

where $f(z) = \cosh(sz)e^{-s^2z^2/2}$. It is a classical fact that $0 \leq \cosh(x) \leq e^{x^2/2}$ for all $x \in \mathbb{R}$ (just compare the Taylor expansions); this implies that $0 \leq f(\lambda) \leq 1$ for all eigenvalues of A_j. Using spectral mapping (4), we see that:

$$\mathrm{spec} f(A_j) = f(\mathrm{spec}(A_j)) \subset [0,1],$$

which implies that $f(A_j) \preceq I$. This proves (13) in this case and finishes the proof of (12) and of the Lemma. □

3.1 Remarks on the original AW approach

A direct adaptation of the original argument of Ahlswede and Winter [2] would lead to an inequality of the form:

$$\mathbb{E}\left[\mathrm{Tr}(e^{sZ_n})\right] \leq \mathrm{Tr}\left(\mathbb{E}\left[e^{s\epsilon_n A_n}\right]\mathbb{E}\left[e^{sZ_{n-1}}\right]\right).$$

One sees that:

$$\mathbb{E}\left[e^{s\epsilon_n A_n}\right] \preceq e^{\frac{s^2A_n^2}{2}} \preceq e^{\frac{s^2\|A_n^2\|}{2}} I.$$

However, only the second inequality seems to be useful, as there is no obvious relationship between

$$\mathrm{Tr}\left(e^{\frac{s^2A_n^2}{2}}\mathbb{E}\left[e^{sZ_{n-1}}\right]\right)$$

and

$$\mathrm{Tr}\left(\mathbb{E}\left[e^{s\epsilon_{n-1}A_{n-1}}\right]\mathbb{E}\left[e^{sZ_{n-2}+\frac{s^2A_n^2}{2}}\right]\right),$$

which is what we would need to proceed with induction. [Note that Golden-Thompson (3) cannot be undone and fails for three summands, [14].] The best one can do with the second inequality is:

$$\mathbb{E}\left[\mathrm{Tr}(e^{sZ_n})\right] \leq d\, e^{\frac{s^2 \sum_{i=1}^n \|A_i\|^2}{2}}.$$

This would give a version of Theorem 1 with $\sum_{i=1}^n \|A_i\|^2$ replacing $\|\sum_{i=1}^n A_i^2\|$. This modified result is always worse than the actual Theorem, and can be dramatically so. For instance, consider the case of a *Wigner matrix* where:

$$Z_n \equiv \sum_{1\leq i\leq j\leq m} \epsilon_{ij} A_{ij}$$

with the ϵ_{ij} i.i.d. standard Gaussian and each A_{ij} has ones at positions (i,j) and (j,i) and zeros elsewhere (we take $d=m$ and $n=\binom{m}{2}$ in this case). Direct calculation reveals:

$$\left\|\sum_{ij} A_{ij}^2\right\| = \|(m-1)I\| = m-1 \ll \binom{m}{2} = \sum_{ij} \|A_{ij}\|^2.$$

We note in passing that neither approach is sharp in this case, as $\|\sum_{ij} \epsilon_{ij} A_{ij}\|$ concentrates around $2\sqrt{m}$ [4].

4 Concentration for rank-one operators

In this section we prove Lemma 1.
Proof: [of Lemma 1] Let

$$\phi(s) \equiv \mathbb{E}\left[\exp\left(s\left\|\frac{1}{n}\sum_{i=1}^n Y_i Y_i^* - \mathbb{E}\left[Y_1 Y_1^*\right]\right\|\right)\right].$$

We will show below that:

$$\forall s\geq 0,\ \phi(s) \leq 2n\, e^{2M^2 s^2/n} \phi(2M^2 s^2/n). \tag{14}$$

By Jensen's inequality, $\phi(2Ms^2/n) \leq \phi(s)^{2M^2 s/n}$ whenever $2M^2 s/n \leq 1$, hence (14) implies:

$$\forall 0\leq s\leq n/2M^2,\ \phi(s) \leq (2n)^{\frac{1}{1-2M^2 s/n}} e^{\frac{2M^2 s^2}{n-2M^2 s}}.$$

Since

$$\forall s\geq 0,\ \mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n Y_i Y_i^* - \mathbb{E}\left[Y_1 Y_1^*\right]\right\| \geq t\right) \leq e^{-st}\phi(s),$$

the Lemma then follows from the choice

$$s \equiv \frac{tn}{8M^2 + 4M^2 t}$$

and a few simple calculations. [Notice that $2M^2 s/n \leq 1/2$ with this choice, hence $1/(1 - 2M^2 s/n) \leq 2$.]

To prove (14), we begin with symmetrization (see e.g. [8]):

$$\phi(s) \leq \mathbb{E}\left[\exp\left(2s\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i Y_i Y_i^*\right\|\right)\right],$$

where $\{\epsilon_i\}_{i=1}^n$ is a Rademacher sequence independent of $Y_1, \ldots, Y_n$. Let $\mathcal{S}$ be the (random) span of $Y_1, \ldots, Y_n$ and $\mathrm{Tr}_{\mathcal{S}}$ denote the trace operation on linear operators mapping $\mathcal{S}$ to itself. Following the argument in Theorem 1, we notice that:

$$\mathbb{E}\left[\exp\left(2s\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i Y_i Y_i^*\right\|\right) \mid Y_1, \ldots, Y_n\right] \leq 2\mathbb{E}\left[\mathrm{Tr}_{\mathcal{S}}\left\{\exp\left(\frac{2s}{n}\sum_{i=1}^n \epsilon_i Y_i Y_i^*\right)\right\} \mid Y_1, \ldots, Y_n\right].$$

Lemma 2 implies:

$$\begin{aligned}\mathbb{E}\left[\exp\left(2s\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i Y_i Y_i^*\right\|\right) \mid Y_1, \ldots, Y_n\right] &\leq 2\mathrm{Tr}_{\mathcal{S}}\left\{\exp\left(\frac{2s^2}{n^2}\sum_{i=1}^n (Y_i Y_i^*)^2\right)\right\} \\ &\leq 2n\exp\left(\left\|\frac{2s^2}{n^2}\sum_{i=1}^n (Y_i Y_i^*)^2\right\|\right),\end{aligned}$$

using spectral mapping (4), the equality "trace = sum of eigenvalues" and the fact that $\mathcal{S}$ has dimension $\leq n$. A quick calculation shows that $0 \preceq (Y_i Y_i^*)^2 = |Y_i|^2 \, Y_i Y_i^* \preceq M^2 Y_i Y_i^*$, hence (5) implies:

$$0 \preceq \frac{2s^2}{n^2}\sum_{i=1}^n (Y_i Y_i^*)^2 \preceq \frac{2M^2 s^2}{n}\left(\frac{1}{n}\sum_{i=1}^n Y_i Y_i^*\right).$$

Therefore:

$$\left\|\frac{2s^2}{n^2}\sum_{i=1}^n (Y_i Y_i^*)^2\right\| \leq \frac{2M^2 s^2}{n}\left\|\frac{1}{n}\sum_{i=1}^n Y_i Y_i^*\right\| \leq \frac{2M^2 s^2}{n}\left\|\frac{1}{n}\sum_{i=1}^n Y_i Y_i^* - \mathbb{E}\left[Y_1 Y_1^*\right]\right\| + \frac{2M^2 s^2}{n}.$$

[We used $\|\mathbb{E}\left[Y_1 Y_1^*\right]\| \leq 1$ in the last inequality.] Plugging this into the conditional expectation above and integrating, we obtain (14):

$$\phi(s) \leq 2n\,\mathbb{E}\left[\exp\left(\frac{2M^2 s^2}{n}\left\|\frac{1}{n}\sum_{i=1}^n Y_i Y_i^* - \mathbb{E}\left[Y_1 Y_1^*\right]\right\| + \frac{2M^2 s^2}{n}\right)\right] = 2ne^{2M^2 s^2/n}\,\phi(2M^2 s^2/n).$$

□

5 Proof sketch for Golden-Thompson inequality

As promised in the Introduction, we sketch an elementary proof of inequality (3). We will need the *Trotter-Lie formula*, a simple consequence of the Taylor formula for e^X:

$$\forall A, B \in \mathbb{C}^{d\times d}_{\mathrm{Herm}}, \lim_{n\to+\infty} (e^{A/n} e^{B/n})^n = e^{A+B}. \tag{15}$$

The second ingredient is the inequality:

$$\forall k \in \mathbb{N}, \forall X, Y \in \mathbb{C}^{d\times d}_{\mathrm{Herm}} : X, Y \succeq 0 \Rightarrow \mathrm{Tr}((XY)^{2^{k+1}}) \leq \mathrm{Tr}((X^2Y^2)^{2^k}). \tag{16}$$

This is proven in of [5] via an argument using the existence of positive-semidefinite square-roots for positive-semidefinite matrices, and the Cauchy-Schwartz inequality for the standard inner product over $\mathbb{C}^{d\times d}$. Iterating (16) implies:

$$\forall X, Y \in \mathbb{C}^{d\times d}_{\mathrm{Herm}} : X, Y \succeq 0 \Rightarrow \mathrm{Tr}((XY)^{2^k}) \leq \mathrm{Tr}(X^{2^k} Y^{2^k}).$$

Apply this to $X = e^{A/2^k}$ and $Y = e^{B/2^k}$ with $A, B \in \mathbb{C}^{d\times d}_{\mathrm{Herm}}$. Spectral mapping (4) implies $X, Y \succeq 0$ and we deduce:

$$\mathrm{Tr}((e^{A/2^k} e^{B/2^k})^{2^k}) \leq \mathrm{Tr}(e^A e^B).$$

Inequality (3) follows from letting $k \to +\infty$, using (15) and noticing that $\mathrm{Tr}(\cdot)$ is continuous.

References

[1] Radoslaw Adamczak, Alexander E. Litvak, Alain Pajor and Nicole Tomczak-Jaegermann. "Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles". Journal of the American Mathematical Society, to appear (2009).

[2] Rudolf Ahlswede and Andreas Winter. "Strong converse for identification via quantum channels." IEEE Transactions on Information Theory 48(3): 569-579 (2002).

[3] Emmanuel Candès and Justin Romberg. "Sparsity and incoherence in compressive sampling." Inverse Problems 23:969985 (2007).

[4] Zoltan Füredi and János Komlós. "The eigenvalues of random symmetric matrices." Combinatorica 1(3), 233-241 (1981).

[5] Sidney Golden. "Lower Bounds for the Helmholtz Function." Physical Review 137, B1127B1128 (1965).

[6] Nathan Halko, Per-Gunnar Martinsson and Joel A. Tropp. "Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions." arXiv:0909.4061 (2009).

[7] Zeph Landau and Alexander Russell. "Random Cayley graphs are expanders: a simplified proof of the Alon-Roichman theorem." The Electronic Journal of Combinatorics, 11(1) (2004).

[8] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces.* Springer (1991).

[9] Françoise Lust-Piquard and Gilles Pisier. "Non commutative Khintchine and Paley inequalities." Arkiv för Matematik, 29(2): 241-260 (1991).

[10] Shahar Mendelson and Alain Pajor. "On singular values of matrices with independent rows". Bernoulli 12(5): 761-773 (2006).

[11] Mark Rudelson. "Random vectors in the isotropic position." Journal of Functional Analysis, 164 (1): 60-72 (1999).

[12] Mark Rudelson and Roman Vershynin. "Sampling from large matrices: an approach through geometric functional analysis." Journal of the ACM 54(4):Article 21 (2007).

[13] Daniel Spielman and Nikhil Srivastava. "Graph sparsification by effective resistances." In Proceedings of the 40th annual ACM symposium on Theory of Computing (STOC 2008).

[14] C.J. Thompson. "Inequality with applications in statistical mechanics." Journal of Mathematical Physics, 6: 1812-1813 (1965).

[15] Joel A. Tropp. "On the conditioning of random subdictionaries." Applied and Computational Harmonic Analysis 25(1): 1-24 (2008).

[16] Roman Vershynin. "Frame expansions with erasures: an approach through the non-commutative operator theory." Applied and Computational Harmonic Analysis 18: 167-176 (2005).

[17] Roman Vershynin. "Spectral norm of products of random and deterministic matrices." Manuscript available from:
`http://www-personal.umich.edu/~romanv/papers/papers.html`.