

LSV-Based Tail Inequalities for Sums of Random Matrices

Chao Zhang

chao.zhang@dlut.edu.cn

Lei Du

dulei@dlut.edu.cn

School of Mathematical Sciences, Dalian University of Technology, Dalian, Liaoning, 116024, P.R.C.

Dacheng Tao

dacheng.tao@uts.edu.au

Centre for Artificial Intelligence, Faculty of Engineering and Information Technology, University of Technology Sydney Ultimo, NSW 2007, Australia.

The techniques of random matrices have played an important role in many machine learning models. In this letter, we present a new method to study the tail inequalities for sums of random matrices. Different from other work (Ahlsvede & Winter, 2002; Tropp, 2012; Hsu, Kakade, & Zhang, 2012), our tail results are based on the largest singular value (LSV) and independent of the matrix dimension. Since the LSV operation and the expectation are noncommutative, we introduce a diagonalization method to convert the LSV operation into the trace operation of an infinitely dimensional diagonal matrix. In this way, we obtain another version of Laplace-transform bounds and then achieve the LSV-based tail inequalities for sums of random matrices.

1 Introduction ---

Eigenproblems play an essential role in many machine learning models—for example, principal component analysis (PCA), Fisher’s linear discriminant analysis (LDA), and spectral clustering. There are also some learning models that are strongly related to nonlinear eigenproblems—for example, 1-spectral clustering and sparse PCA (Hein & Bühler, 2010), and balanced graph cuts and RatioDCA-Prox (Hein & Setzer, 2011). We refer to Jost, Setzer, and Hein (2014) for details.

In general, since the learning data are assumed to be drawn from probability distributions, the eigenproblem of random matrices naturally becomes one of the most important concerns in the mathematical foundations of machine learning. For example, Bian and Tao (2014) use random matrix tools to study generalization bounds of Fisher’s linear discriminant analysis. Ahlsvede and Winter (2002) developed the large deviation inequalities

for sums of random variables taking values in self-adjoint operators. Vershynin (2010) analyzed the lower and the upper bounds of the singular values of a random matrix $\mathbf{A}_{m \times n}$ when the dimension m (or n) goes to the infinity. Tropp (2012) provided a powerful and user-friendly framework to obtain tail bounds for the extreme eigenvalues of sums of self-adjoint random matrices. Different from the explicit matrix dimension appearing in Tropp (2012), Hsu, Kakade, and Zhang (2012) developed exponential tail inequalities for sums of real, symmetric, random matrices that depend only on intrinsic dimensions. By the method of exchangeable pairs, Mackey, Jordan, Chen, Farrell, and Tropp (2014) present concentration inequalities for the extreme eigenvalues of sums of Hermitian random matrices.

1.1 Eigenproblems. For a number $\lambda \in \mathbb{C}$, let $\mathbf{F}(\lambda) : \mathbb{C} \rightarrow \mathbb{C}^{d \times d}$ be a matrix-valued function and $\mathbf{x} \in \mathbb{C}^n$ be a vector. Some kinds of eigenproblems can be derived from the equation

$$\mathbf{F}(\lambda)\mathbf{x} = 0. \quad (1.1)$$

For example, setting $\mathbf{F}(\lambda) = (\mathbf{A} - \lambda\mathbf{I})$ with $\mathbf{A} \in \mathbb{C}^{d \times d}$ makes equation 1.1 the standard eigenproblem,

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}, \quad (1.2)$$

where the vector $\mathbf{x} \in \mathbb{R}^d$ is the eigenvector corresponding to the eigenvalue λ . The standard eigenproblem, equation 1.2, plays an essential role in PCA and spectral clustering.

If $\mathbf{F}(\lambda) = (\mathbf{A} - \lambda\mathbf{B})$ with $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{d \times d}$, equation 1.1 is the so-called generalized eigenproblem,

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{B}\mathbf{x}, \quad (1.3)$$

whose solutions λ are called the generalized eigenvalues of \mathbf{A} with regard to \mathbf{B} . Especially if matrix \mathbf{B} is nonsingular, the generalized eigenproblem, equation 2.3, can be transformed into a standard eigenproblem, $\mathbf{B}^{-1}\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$, which provides solutions to Fisher's LDA. Furthermore, when $\mathbf{F}(\lambda)$ is a nonlinear function, equation 1.1 refers to so-called nonlinear eigenproblems, which are strongly related to some learning models (see Hein & Bühler, 2010; Hein & Setzer, 2011; Jost et al., 2014).

The eigenproblem is also strongly related to many research interests of dynamic systems. For example, the sign of the largest eigenvalue predicts persistence or extinction of models in spatial ecology, and its value describes the dependence of models in spatial ecology on the geometry and size of the underlying habitat region (Cantrell & Cosner, 2004). In online social networks, the largest eigenvalue is the bifurcation point of information

spreading or vanishing, and thus it predicts the distance of information persistence or extinction (Dai, Ma, Wang, Wang, & Xu, 2015). In the field of theoretical neuroscience, the synaptic connections of a neuronal network can be modeled as a random matrix whose entries are the strengths of synapses between all pairs of neurons and obey the appropriate distributions (e.g., gaussian). The eigenproblem of the random matrix has been applied to reveal the relationship between connectivity and dynamics in neuronal networks (Rajan & Abbott, 2006; Muir & Mrcsic-Flogel, 2015).

1.2 Background and Motivation. This section summarizes some concerns in existing work on tail inequalities for sums of random matrices; it is these concerns that motivate this letter as well.

As shown in proposition 3.1 of Tropp (2012), the Laplace-transform bound

$$\Pr\{\lambda_{\max}(\mathbf{Y}) \geq t\} \leq \inf_{\theta > 0} \{e^{-\theta t} \cdot \mathbb{E} \operatorname{tr} e^{\theta \mathbf{Y}}\}$$

provides a starting point to study tail inequalities for sums of random matrices. The key to tail inequalities is to bound the term $\mathbb{E} \operatorname{tr} e^{\theta \mathbf{Y}}$ with $\mathbf{Y} = \sum_{k=1}^K \mathbf{X}_k$, where $\mathbf{X}_1, \dots, \mathbf{X}_K$ are independent, random, Hermitian matrices. In the literature, there are mainly two types of bounds: one is given by Ahlswede and Winter (2002),

$$\begin{aligned} \mathbb{E} \operatorname{tr} e^{\theta \mathbf{Y}} &= \operatorname{tr} \mathbb{E} e^{\theta \mathbf{Y}} \leq (\operatorname{tr} \mathbf{I}) \cdot \left[\prod_k \lambda_{\max}(\mathbb{E} e^{\theta \mathbf{X}_k}) \right] \\ &= d \cdot \exp \left(\sum_k \lambda_{\max}(\log \mathbb{E} e^{\theta \mathbf{X}_k}) \right), \end{aligned} \tag{1.4}$$

and the other is presented by Tropp (2012):

$$\mathbb{E} \operatorname{tr} e^{\theta \mathbf{Y}} \leq d \cdot \exp \left(\lambda_{\max} \left(\sum_k \log \mathbb{E} e^{\theta \mathbf{X}_k} \right) \right). \tag{1.5}$$

As shown above, Ahlswede-Winter’s bound, equation 1.4, and Tropp’s bound, equation 1.5, are both based on the matrix trace, which brings convenience to obtaining tail inequalities.¹

¹For example, the fact that $\mathbb{E} \operatorname{tr}(\mathbf{A}) = \operatorname{tr}(\mathbb{E} \mathbf{A})$ makes the final tail inequalities valid under a milder condition: the expectation of the random terms can be bounded by some deterministic terms instead of the condition that all the random terms are bounded.

However, the trace operation inevitably makes the bounds 1.4 and 1.5 loose. More specifically, it is the reason that the two bounds are dependent on the matrix dimension d . It is noteworthy that for real symmetric matrices, Hsu et al. (2012) present the bounds dependent on the intrinsic dimension $\text{tr}(\mathbf{Y})/\lambda_{\max}(\mathbf{Y}) \leq d$ in the setting of real symmetric matrices. Since these results are dependent on the matrix dimension d , they may be more suitable to a scenario of low-dimensional matrices.

One of main contributions in Tropp's (2012) framework is to apply Lieb's concavity theorem to achieve the right-hand side of equation 1.5:

$$d \cdot \exp \left(\lambda_{\max} \left(\sum_k \log E e^{\theta \mathbf{X}_k} \right) \right)$$

instead of the right-hand side of equation 1.4,

$$d \cdot \exp \left(\sum_k \lambda_{\max}(\log E e^{\theta \mathbf{X}_k}) \right),$$

which is derived from Golden-Thompson trace inequality. Note that the former is tighter than the latter because of the fact that $\lambda_{\max}(\sum_k \mathbf{X}_k) \leq \sum_k \lambda_{\max}(\mathbf{X}_k)$.

1.3 Overview of Main Results. Being motivated by the concerns we have noted, this letter presents new tail inequalities for sums of random matrices, which are based on the LSV operation σ_1 and independent of the matrix dimension d .²

We first present Laplace-transform bounds based on the LSV operation $\sigma_1(\mathbf{X})$. As addressed in remark 1, since the LSV operation does not hold the associative law for addition, it is difficult to obtain the results with the term $\sigma_1(\sum_k \mathbf{X}_k)$. To overcome this limitation, we attempt to apply a diagonalization method (DM) to convert the operation σ_1 into the trace of an infinitely dimensional diagonal matrix. We then develop the DM-based Laplace-transform bounds and present the relevant tail inequalities for the spectral radii of sums of random matrices.

Compared with previous work (Ahlsvede & Winter, 2002; Tropp, 2012; Hsu et al., 2012), our resulting inequalities are independent of the matrix dimension d and thus suitable to the case of high-dimensional matrices. As addressed in remark 4, however, there are two things to be noted: our tail inequalities may not be applicable to sums of a large quantity of random matrices and the results in equations 3.16 and 3.17 are affected by the

²Given a matrix \mathbf{A} , the LSV $\sigma_1(\mathbf{A})$ actually is the spectral norm of \mathbf{A} , that is, the negative square root of the maximum eigenvalue of $\mathbf{A}^H \mathbf{A}$ (see Higham, 2008).

choice of g_0 and $\{\mathbf{A}_k\}$. Therefore, if there are wiser choices of g_0 and $\{\mathbf{A}_k\}$, the relevant tail inequalities may overcome the aforementioned obstacles.

1.4 Organization of the Letter. The rest of this letter is organized as follows. Section 2 presents the LSV-based Laplace-transform inequalities, and the DM-based results are given in section 3. The last section concludes the letter, and some proofs are in the appendix.

2 LSV-Based Laplace-Transform Inequalities _____

We first give some preliminaries on the largest singular value (LSV):

Lemma 1. *Given two matrices A and B , there holds that*

- i. $\sigma_1(A \cdot B) \leq \sigma_1(A) \cdot \sigma_1(B)$.
- ii. $\sigma_1(A + B) \leq \sigma_1(A) + \sigma_1(B)$.

Proof. Inequality i is a special case of theorem H.1.c in Marshall, Olkin, and Arnold (2010). Inequality ii holds because of the triangle inequality of the spectral norm.

As Tropp (2012) addressed, Laplace-transform bounds provide a starting point for obtaining the tail inequalities for the sum of random matrices. The relevant results in existing work are built for the largest eigenvalues (see Ahlswede & Winter, 2002; Tropp, 2012), while this letter considers LSV-based Laplace-transform bounds for random matrices.

Theorem 1. *Let $\{\mathbf{X}_k\}_{k=1}^K$ be a sequence of independent, square random matrices. Then there holds that for any $t > 0$,*

$$\begin{aligned} \Pr \left\{ \sigma_1 \left(\sum_{k=1}^K \mathbf{X}_k \right) \geq t \right\} &\leq \inf_{\theta > 0} \left\{ e^{-\theta t} \cdot \prod_k E e^{\sigma_1(\theta \mathbf{X}_k)} \right\} \\ &\leq \inf_{\theta > 0} \left\{ e^{-\theta t} \cdot \exp \left(K \cdot \log \left(\frac{1}{K} \sum_k E e^{\sigma_1(\theta \mathbf{X}_k)} \right) \right) \right\}. \end{aligned} \tag{2.1}$$

Proof. According to Markov’s inequality and lemma 1, we have

$$\begin{aligned} \Pr \left\{ \sigma_1 \left(\sum_k \mathbf{X}_k \right) \geq t \right\} &= \Pr \left\{ \sigma_1 \left(\theta \sum_k \mathbf{X}_k \right) \geq \theta t \right\} \\ &= \Pr \left\{ e^{\sigma_1(\theta \sum_k \mathbf{X}_k)} \geq e^{\theta t} \right\} \leq e^{-\theta t} \cdot E e^{\sigma_1(\theta \sum_k \mathbf{X}_k)} \tag{*} \\ &\leq e^{-\theta t} \cdot E e^{\sum_k \sigma_1(\theta \mathbf{X}_k)} = e^{-\theta t} \cdot \prod_k E e^{\sigma_1(\theta \mathbf{X}_k)} \end{aligned}$$

$$\begin{aligned}
&= e^{-\theta t} \cdot \exp\left(\sum_k \log E e^{\sigma_1(\theta X_k)}\right) \\
&\leq e^{-\theta t} \cdot \exp\left(K \cdot \log\left(\frac{1}{K} \sum_k E e^{\sigma_1(\theta X_k)}\right)\right), \tag{2.2}
\end{aligned}$$

which holds for any $\theta > 0$ and the step labeled * holds because of the independence of X_k . Taking an infimum completes the proof.

Remark 1. Compared with the previous results, equations 1.4 and 1.5, the absence of the trace operation makes the upper bound of $\Pr\{\sigma_1(\sum_k X_k) \geq t\}$ independent of the matrix dimension. In contrast with Tropp's bound, equation 1.5, since the LSV operation and the expectation are noncommutative, it is difficult to obtain the Laplace-transform bound with the term $\sigma_1(\sum_k X_k)$.

To handle this issue, section 3 shows the tail inequalities incorporating the term $\sigma_1(\sum_k A_k)$, where A_k are the fixed matrices dominating the behaviors of random matrices X_k ($1 \leq k \leq K$).

3 Diagonalization Method and Tail Inequalities

In this section, we introduce the diagonalization method (DM) to convert the LSV operation to a more convenient form for the discussion that follows. Then we present the DM-based Laplace-transform bounds for sums of random matrices, as well as the relevant tail inequalities.

3.1 Diagonalization Method. The trace operation supports the fact $E \operatorname{tr}(\mathbf{A}) = \operatorname{tr}(E \mathbf{A})$; in contrast, it is followed from Jensen's inequality that $E \sigma_1(\mathbf{A}) \geq \sigma_1(E \mathbf{A})$. To overcome this limitation of LSV, we propose a method to convert the LSV operation into the trace of an infinitely dimensional diagonal matrix with the entries being functions of LSV. Compared to the LSV operation, the trace form has better operational properties; for example, Lieb's concavity theorem becomes valid in this setting, and thus we can obtain tighter tail inequalities.

The following shows the diagonalization method to realize the LSV operation by using the trace operation.

Proposition 1. *Given a matrix $X \in \mathbb{C}^{d \times d}$, there holds that for any $\theta > 0$,*

$$E e^{\sigma_1(\theta X)} = e^{-1} \cdot \operatorname{tr}(E e^{D[\theta; X]}), \tag{3.1}$$

where

$$D[\theta; \mathbf{X}] := \mathbf{\Lambda} \left\{ 0, 0, \log \frac{1}{2!}, \log \frac{1}{3!}, \dots \right\} \tag{3.2}$$

$$+ \mathbf{\Lambda} \left\{ 0, \log (\theta \sigma_1(\mathbf{X}) + 1), 2 \log (\theta \sigma_1(\mathbf{X}) + 1), 3 \log (\theta \sigma_1(\mathbf{X}) + 1), \dots \right\}$$

and $\mathbf{\Lambda}\{\dots\}$ stands for the diagonal matrix.

Note that the result, equation 3.1, is obtained by using Taylor’s expansion of e^x to convert the LSV operation $\sigma_1(\theta \mathbf{X})$ into the trace operation for an infinitely dimensional diagonal matrix $\mathbf{D}[\theta; \mathbf{X}]$ (see appendix A). Subsequently, we use the digitalization method to obtain an upper bound of $E e^{\theta \sigma_1(\sum_k \mathbf{X}_k)}$.

Proposition 2 (subadditivity). *Let $\{\mathbf{X}_1, \dots, \mathbf{X}_K\}$ be a sequence of independent random matrices. Then there holds that for any $\theta > 0$,*

$$E e^{\theta \sigma_1(\sum_{k=1}^K \mathbf{X}_k)} \leq e^{-K} \cdot \text{tr} \exp \left(\sum_{k=1}^K \log E e^{\mathbf{D}[\theta; \mathbf{X}_k]} \right), \tag{3.3}$$

where for any $1 \leq k \leq K$,

$$D[\theta; \mathbf{X}_k] := \mathbf{\Lambda} \left\{ 0, 0, \log \frac{1}{2!}, \log \frac{1}{3!}, \dots \right\} \tag{3.4}$$

$$+ \mathbf{\Lambda} \left\{ 0, \log (\theta s_k + 1), 2 \log (\theta s_k + 1), 3 \log (\theta s_k + 1), \dots \right\}$$

with $s_k := \sigma_1(\mathbf{X}_k)$.

This result is derived from the subadditivity of the operation $\mathbf{D}[\theta; \mathbf{X}_k]$, that is,

$$\sum_k \mathbf{D}[\theta; \mathbf{X}_k] \geq \mathbf{D} \left[\theta; \sum_k \mathbf{X}_k \right] + (K - 1) \cdot \mathbf{\Lambda} \left\{ 0, 0, \log \frac{1}{2!}, \log \frac{1}{3!}, \dots \right\}.$$

Details are referred to the proof of proposition 2 in appendix A.

3.2 DM-Based Laplace-transform Bounds. Next, we develop another version of Laplace-transform bounds as the starting point for the relevant tail inequalities.

Proposition 3. Given a random matrix $\mathbf{X} \in \mathbb{C}^{d \times d}$, there holds that for any $t > 0$,

$$\Pr\{\sigma_1(\mathbf{X}) \geq t\} \leq e^{-1} \cdot \inf_{\theta > 0} \{e^{-\theta t} \cdot \text{tr}(E e^{\mathbf{D}[\theta; \mathbf{X}]})\}, \quad (3.5)$$

where $\mathbf{D}[\theta; \mathbf{X}]$ is defined in equation 3.2.

Proof. From theorem 2.1, we have

$$\Pr\{\sigma_1(\mathbf{X}) \leq t\} \leq \inf_{\theta > 0} \{e^{-\theta t} \cdot E e^{\sigma_1(\theta \mathbf{X})}\}, \quad (3.6)$$

which is of the case $K = 1$. Then the combination of equation 3.6 and proposition 1 leads to the result, equation 3.5. This completes the proof.

This bound is suitable only to the scenario of a single random matrix. Next, we extend the bound, equation 3.5, to the scenario of sums of random matrices. The following result can be obtained by combining proposition 2 and the second line of equation 2.2.

Theorem 2. Let $\{\mathbf{X}_k\}_{k=1}^K$ be a sequence of independent random matrices. Then there holds that for any $t \in \mathbb{R}$,

$$\begin{aligned} & \Pr\left\{\sigma_1\left(\sum_{k=1}^K \mathbf{X}_k\right) \geq t\right\} \\ & \leq e^{-K} \cdot \inf_{\theta > 0} \left\{e^{-\theta t} \cdot \text{tr} \exp\left(\sum_{k=1}^K \log E e^{\mathbf{D}[\theta; \mathbf{X}_k]}\right)\right\}, \end{aligned} \quad (3.7)$$

where $\mathbf{D}[\theta; \mathbf{X}_k]$ is defined in equation 3.4 for any $1 \leq k \leq K$.

As shown in theorem 2, we can use the diagonalization method to convert the LSV-based inequalities given in theorem 1 into the Laplace-transform bounds based on the trace of an infinitely dimensional diagonal matrix. However, it could be impossible to directly derive the inequalities incorporating the term $\mathbf{D}[\theta; \sum_k \mathbf{X}_k]$ from the inequality 3.7. Therefore, we give the following scheme to handle this issue.

Theorem 3 (superadditivity). Consider a finite sequence $\{\mathbf{X}_k\}_{k=1}^K$ of independent square random matrices. Let $\{\mathbf{B}_k\}_{k=1}^K$ be a sequence of fixed matrices such that for any $1 \leq k \leq K$,

$$E \log(\sigma_1(\theta \mathbf{X}_k) + 1) \leq \log(\sigma_1(\theta \mathbf{B}_k) + 1). \quad (3.8)$$

Assume that there exists a function $g_0(\theta) : (0, \infty) \rightarrow [0, \infty]$ and a sequence $\{\mathbf{A}_k\}_{k=1}^K$ of fixed matrices that satisfy the relation: for any $\theta > 0$,

$$\sum_{k=1}^K D[\theta; \mathbf{B}_k] \leq D\left[g_0(\theta); \sum_k \mathbf{A}_k\right] + (K - 1) \cdot \mathbf{A} \left\{0, 0, \log \frac{1}{2!}, \log \frac{1}{3!}, \dots\right\}. \tag{3.9}$$

Define $\eta := \sigma_1(\sum_k \mathbf{A}_k)$. Then for all $t \in \mathbb{R}$,

$$Pr \left\{ \sigma_1 \left(\sum_k \mathbf{X}_k \right) \geq t \right\} \leq \inf_{\theta > 0} \left\{ e^{-\theta t + g_0(\theta) \cdot \eta} \right\}. \tag{3.10}$$

In theorem 3, we introduce the sequence $\{\mathbf{B}_k\}$ of fixed matrices to control the expectations of random matrices $\{\mathbf{X}_k\}$ (see equation 3.8). Then we use the function $g_0(\theta)$ and the sequence $\{\mathbf{A}_k\}$ of fixed matrices to realize the superadditivity of the operation $D[\cdot; \cdot]$. Therefore, the validity of theorem 3 depends on the existence of the function $g_0(\theta)$ and the sequence $\{\mathbf{A}_k\}$. Next, we take an example to demonstrate the existence of $g_0(\theta)$ and $\{\mathbf{A}_k\}$.

Remark 2 [LSV choice of \mathbf{B}_k]. From equation 3.8, since $\sigma_1(\theta \mathbf{A}) = \theta \sigma_1(\mathbf{A})$ ($\forall \theta > 0$), we have for any $1 \leq k \leq K$,

$$\sigma_1(\mathbf{B}_k) \geq \frac{1}{\theta} \exp(E \log(\sigma_1(\theta \mathbf{X}_k) + 1)) - \frac{1}{\theta}. \tag{3.11}$$

Let $\mathbf{X}_k^{(1)}, \dots, \mathbf{X}_k^{(N)}$ be the i.i.d. observations of the random matrix \mathbf{X}_k , and then the expectation term $E \log(\sigma_1(\theta \mathbf{X}_k) + 1)$ can be approximated by the empirical quantity

$$\frac{1}{N} \sum_{n=1}^N \log(\sigma_1(\theta \mathbf{X}_k^{(n)}) + 1). \tag{3.12}$$

The combination of equations 3.11 and 3.12 leads to

$$\begin{aligned} \sigma_1(\mathbf{B}_k) &\geq \frac{1}{\theta} \exp\left(\frac{1}{N} \sum_{n=1}^N \log(\sigma_1(\theta \mathbf{X}_k^{(n)}) + 1)\right) - \frac{1}{\theta} \\ &= N \sqrt{\prod_{n=1}^N \left(\sigma_1(\mathbf{X}_k^{(n)}) + \frac{1}{\theta}\right)} - \frac{1}{\theta}. \end{aligned} \tag{3.13}$$

It is followed from the fact $N\sqrt{x_1x_2\cdots x_N} \leq \frac{x_1+x_2+\cdots+x_N}{N}$ that

$$N \sqrt{\prod_{n=1}^N \left(\sigma_1(\mathbf{X}_k^{(n)}) + \frac{1}{\theta} \right)} - \frac{1}{\theta} \leq \frac{\sigma_1(\mathbf{X}_k^{(1)}) + \cdots + \sigma_1(\mathbf{X}_k^{(N)})}{N}. \tag{3.14}$$

Therefore, a reasonable LSV choice of \mathbf{B}_k turns out to be

$$\sigma_1(\mathbf{B}_k) = \frac{\sigma_1(\mathbf{X}_k^{(1)}) + \cdots + \sigma_1(\mathbf{X}_k^{(N)})}{N},$$

when the sample number N is sufficiently large.

Remark 3 (existence of $g_0(\theta)$ and $\{\mathbf{A}_k\}$). Let \mathbf{A}_k be the matrix with the largest singular value $\sigma_1(\mathbf{A}_k) = f_K(\sigma_1(\mathbf{B}_k))$ where

$$f_K(s) = \sum_{i=0}^{K-1} \frac{\binom{K-1}{i} \cdot s^{K-i}}{K}.$$

Note that the function f_K actually supports the inequality

$$\log \left(1 + \sum_{k=1}^K f_K(s_k) \right) \geq \sum_{k=1}^K \log(s_k + 1). \tag{3.15}$$

We then realize the superadditivity of $\mathbf{D}[\cdot; \cdot]$ as follows:

$$\sum_k^K \mathbf{D}[\theta; \mathbf{B}_k] \leq \mathbf{D} \left[g_0(\theta); \sum_k^K \mathbf{A}_k \right] + (K-1) \cdot \Lambda \left\{ 0, 0, \log \frac{1}{2!}, \log \frac{1}{3!}, \dots \right\}$$

with $g_0(\theta) \geq \max\{\theta, \theta^K\}$. The details and some special cases are given in appendix B.

In summary, the key to theorem 3 is to develop the function f_K such that inequality 3.15 holds. If such a function f_K is found, the fixed matrix sequence $\{\mathbf{A}_k\}$ and the function $g_0(\theta)$ should satisfy the inequalities $f_K(\sigma_1(\mathbf{B}_k)) \leq \sigma_1(\mathbf{A}_k)$ and $f_K(\theta s) \leq g_0(\theta) \cdot f_K(s)$, respectively.

3.3 DM-Based Tail Inequalities. Based on theorem 3, we can obtain the DM-based tail inequalities for sums of random matrices as follows:

Theorem 4. Assume that $\{\mathbf{X}_k\}_{k=1}^K$ are independent square random matrices. Follow the notations in remark 3, and let $\eta := \sigma_1(\sum_k \mathbf{A}_k)$. Then, there holds that

1. For any $t > 0$,

$$\Pr \left\{ \sigma_1 \left(\sum_{k=1}^K \mathbf{X}_k \right) \geq t \right\} \leq \exp \left(-\eta \cdot h \left(\frac{t}{K\eta} \right) \right) \tag{3.16}$$

with $h(x) := x \log(x) - x$.

2. For any $t > \eta$,

$$\begin{aligned} \Pr \left\{ \sigma_1 \left(\sum_{k=1}^K \mathbf{X}_k \right) \geq t \right\} &\leq \exp \left(-t \left[\frac{1}{K} \left(\frac{t}{\eta} - 1 \right) \right]^{\frac{1}{K-1}} \right. \\ &\left. + \eta \left(\left[\frac{1}{K} \left(\frac{t}{\eta} - 1 \right) \right]^{\frac{1}{K-1}} + \left[\frac{1}{K} \left(\frac{t}{\eta} - 1 \right) \right]^{\frac{K}{K-1}} \right) \right). \end{aligned} \tag{3.17}$$

Proof. Since $\max\{\theta, \theta^K\} \leq e^{K\theta}$, set $g_0(\theta) = e^{K\theta}$ and consider the infimum $\inf_{\theta>0}\{-\theta t + g_0(\theta)\eta\}$. When $\theta = \frac{1}{K} \log(t/K\eta)$, the right-hand side of the equation 3.10 reaches the minimum. Substituting $\theta = \frac{1}{K} \log(t/K\eta)$ into equation 3.10 leads to inequality 3.16. Similarly, letting $g_0(\theta) = \theta + \theta^K$ can achieve the result in equation 3.17.

Remark 4. Note that the superadditivity 3.9 is not quite the ideal result, because \mathbf{A}_k will become much larger than \mathbf{B}_k ($1 \leq k \leq K$) when K is large. It implies that our results, equations 3.16 and 3.17, are not suitable to the case of a large quantity of random matrices. Moreover, the choices of $g_0(\theta)$ and $\{\mathbf{A}_k\}$ are essential to the effectiveness of the tail inequalities given in theorem 3. If we can find wiser choices of $g_0(\theta)$ and $\{\mathbf{A}_k\}$, that obstacle may be overcome.

4 Conclusion

In this letter, we present LSV-based tail inequalities for sums of random matrices. Unlike previous work (Hsu et al., 2012; Tropp, 2012; Ahlswede and Winter, 2002), our results are independent of the matrix-dimension and thus are more applicable to the eigenproblems for sums of high-dimensional random matrices. However, it is noteworthy that they are not suitable to the scenario of quantities of summands as discussed in remark 4. In future work, we will find some wiser choices of $g_0(\theta)$ and $\{\mathbf{A}_k\}$ to overcome the obstacle and use the resulting tail inequalities to analyze the dynamics of neuronal networks.

Appendix A: Proofs of Some Results

Here, we prove propositions 1 and 2.

Proof of Proposition 1. Given a matrix \mathbf{X} and $\theta > 0$, briefly denote $s := \sigma_1(\mathbf{X})$. Then the term $\mathbf{D}[\theta; \mathbf{X}]$ defined in equation 3.4 can be equivalently written as

$$\mathbf{D}[\theta; \mathbf{X}] = \log \Lambda \left\{ 1, \theta s + 1, \frac{(\theta s + 1)^2}{2!}, \frac{(\theta s + 1)^3}{3!}, \dots \right\}. \tag{A.1}$$

It follows from Taylor’s expansion $e^x = 1 + \sum_k \frac{x^k}{k!}$ that

$$e^{\theta s + 1} = \text{tr} \left(\Lambda \left\{ 1, \theta s + 1, \frac{(\theta s + 1)^2}{2!}, \frac{(\theta s + 1)^3}{3!}, \dots \right\} \right). \tag{A.2}$$

The combination of equations A.1 and A.2 leads to

$$e^{\sigma_1(\theta \mathbf{X})} = e^{\theta \sigma_1(\mathbf{X})} = e^{-1} \cdot e^{\theta \sigma_1(\mathbf{X}) + 1} = e^{-1} \cdot \text{tr}(e^{\mathbf{D}[\theta; \mathbf{X}]})$$

Proof of Proposition 2. Set $\mathbf{Y} = \sum_{k=1}^K \mathbf{X}_k$, and there holds that

$$\sigma_1(\theta \mathbf{Y}) = \sigma_1 \left(\theta \sum_k \mathbf{X}_k \right) \leq \sum_k \sigma_1(\theta \mathbf{X}_k). \tag{A.3}$$

Thus, we have

$$\begin{aligned} \mathbf{D}[\theta; \mathbf{Y}] &= \Lambda \left\{ 0, \log(\sigma_1(\theta \mathbf{Y}) + 1), 2 \log(\sigma_1(\theta \mathbf{Y}) + 1), \dots \right\} \\ &\quad + \Lambda \left\{ 0, 0, \log \frac{1}{2!}, \log \frac{1}{3!}, \dots \right\} \\ &\leq \sum_{k=1}^K \Lambda \left\{ 0, \log(\sigma_1(\theta \mathbf{X}_k) + 1), 2 \log(\sigma_1(\theta \mathbf{X}_k) + 1), \dots \right\} \\ &\quad + \Lambda \left\{ 0, 0, \log \frac{1}{2!}, \log \frac{1}{3!}, \dots \right\} \\ &= \left(\sum_{k=1}^K \mathbf{D}[\theta; \mathbf{X}_k] \right) - (K - 1) \cdot \Lambda \left\{ 0, 0, \log \frac{1}{2!}, \log \frac{1}{3!}, \dots \right\}, \tag{A.4} \end{aligned}$$

where the inequality is also followed from the fact that

$$\log(x + y + 1) \leq \log(x + 1) + \log(y + 1), \quad \forall x, y > 0.$$

It is noteworthy that if \mathbf{A} and \mathbf{B} are both diagonal matrices, then $e^{\mathbf{A}+\mathbf{B}} = e^{\mathbf{A}} \cdot e^{\mathbf{B}}$ (see Higham, 2008, theorem 10.2). By equation A.4, we then have

$$\begin{aligned} \text{tr } e^{\mathbf{D}[\theta; \mathbf{Y}]} &\leq \text{tr} \left(\exp \left((1-K) \cdot \mathbf{\Lambda} \left\{ 0, 0, \log \frac{1}{2!}, \log \frac{1}{3!}, \dots \right\} + \sum_{k=1}^K \mathbf{D}[\theta; \mathbf{X}_k] \right) \right) \\ &= \text{tr} \left(\left(\mathbf{\Lambda} \left\{ 1, 1, \frac{1}{2!}, \frac{1}{3!}, \dots \right\} \right)^{1-K} \cdot \exp \left(\sum_{k=1}^K \mathbf{D}[\theta; \mathbf{X}_k] \right) \right) \\ &\leq \left(\text{tr } \mathbf{\Lambda} \left\{ 1, 1, \frac{1}{2!}, \frac{1}{3!}, \dots \right\} \right)^{1-K} \cdot \text{tr} \left(\exp \left(\sum_{k=1}^K \mathbf{D}[\theta; \mathbf{X}_k] \right) \right) \\ &= e^{1-K} \cdot \text{tr} \left(\exp \left(\sum_{k=1}^K \mathbf{D}[\theta; \mathbf{X}_k] \right) \right), \end{aligned} \tag{A.5}$$

where the second inequality is followed from the fact that $\text{tr}(\mathbf{A} \cdot \mathbf{B}) \leq \text{tr}(\mathbf{A}) \cdot \text{tr}(\mathbf{B})$ if \mathbf{A} and \mathbf{B} are positive definite diagonal matrices. According to lemma 3.4 of Tropp (2012), it holds that

$$\text{E tr} \left(\exp \left(\sum_{k=1}^K \mathbf{D}[\theta; \mathbf{X}_k] \right) \right) \leq \text{tr} \left(\exp \left(\sum_{k=1}^K \log \text{E } e^{\mathbf{D}[\theta; \mathbf{X}_k]} \right) \right). \tag{A.6}$$

By combining equations A.5 and A.6 and proposition 1, we finally obtain

$$\begin{aligned} \text{E } e^{\sigma_1(\theta \sum_k \mathbf{X}_k)} &= e^{-1} \cdot \text{E } e^{\sigma_1(\theta \sum_k \mathbf{X}_k)+1} \\ &= e^{-1} \cdot \text{E tr } e^{\mathbf{D}[\theta; \sum_k \mathbf{X}_k]} \\ &\leq e^{-K} \cdot \text{tr} \exp \left(\sum_{k=1}^K \log \text{E } e^{\mathbf{D}[\theta; \mathbf{X}_k]} \right). \end{aligned}$$

Appendix B: An Example of Choosing $g_0(\theta)$ and \mathbf{A}_k _____

When $K = 2$, since $s_1 \cdot s_2 \leq (s_1^2 + s_2^2)/2$,

$$\begin{aligned} \log(1 + s_1) + \log(1 + s_2) &= \log(1 + s_1)(1 + s_2) \\ &= \log(1 + s_1 + s_2 + s_1 \cdot s_2) \\ &\leq \log \left(1 + s_1 + s_2 + \frac{(s_1^2 + s_2^2)}{2} \right) \\ &= \log \left(1 + \left(s_1 + \frac{s_1^2}{2} \right) + \left(s_2 + \frac{s_2^2}{2} \right) \right), \end{aligned}$$

which implies that $f_2(s) = s + \frac{s^2}{2}$.

When $K = 3$, since $s_1 \cdot s_2 \cdot s_3 \leq (s_1^3 + s_2^3 + s_3^3)/3$,

$$\begin{aligned} & \log(1 + s_1) + \log(1 + s_2) + \log(1 + s_3) \\ &= \log(1 + s_1)(1 + s_2)(1 + s_3) \\ &= \log\left(1 + s_1 + s_2 + s_3 + s_1s_2 + s_1s_3 + s_2s_3 + s_1s_2s_3\right) \\ &\leq \log\left(1 + s_1 + s_2 + s_3 + \frac{(s_1^2 + s_2^2)}{2} + \frac{(s_1^2 + s_3^2)}{2}\right. \\ &\quad \left.+ \frac{(s_2^2 + s_3^2)}{2} + \frac{(s_1^3 + s_2^3 + s_3^3)}{3}\right) \\ &= \log\left(1 + \left(s_1 + s_1^2 + \frac{s_1^3}{3}\right) + \left(s_2 + s_2^2 + \frac{s_2^3}{3}\right) + \left(s_3 + s_3^2 + \frac{s_3^3}{3}\right)\right), \end{aligned}$$

which implies that $f_3(s) = s + s^2 + \frac{s^3}{3}$.

For the general case K , denote Ω_k as a set of cardinality k . Since $\prod_{k=1}^K s_k \leq \frac{\sum_{k=1}^K s_k^k}{K}$, we have

$$\begin{aligned} & \sum_{k=1}^K \log(1 + s_k) = \log\left(\prod_{k=1}^K (1 + s_k)\right) \\ &= \log\left(1 + \sum_{k=1}^K \left(\sum_{\Omega_k \subset \{1, \dots, K\}} \prod_{i \in \Omega_k} s_i\right)\right) \\ &\leq \log\left(1 + \sum_{k=1}^K \left(\sum_{\Omega_k \subset \{1, \dots, K\}} \left(\sum_{i \in \Omega_k} \frac{s_i^k}{k}\right)\right)\right) \\ &\leq \log\left(1 + \sum_{k=1}^K \left(\sum_{i=0}^{K-1} \frac{\binom{K-1}{K-1-i} \cdot s_k^{K-i}}{K}\right)\right) \\ &= \log\left(1 + \sum_{k=1}^K f_K(s_k)\right) \end{aligned}$$

with

$$f_K(s) = \sum_{i=0}^{K-1} \frac{\binom{K-1}{K-1-i} \cdot s^{K-i}}{K} = \sum_{i=0}^{K-1} \frac{\binom{K-1}{i} \cdot s^{K-i}}{K}.$$

On the other hand, the expression of f_K can also lead to

$$f_K(\theta s) = \sum_{i=0}^{K-1} \frac{\binom{K-1}{i} \cdot (\theta s)^{K-i}}{K} \leq \begin{cases} \theta \cdot f_K(s), & \text{if } 0 < \theta \leq 1; \\ \theta^K \cdot f_K(s), & \text{if } \theta > 1, \end{cases}$$

which implies that $f_K(\theta s) \leq g_0(\theta) \cdot f_K(s)$ if the function $g_0(\theta) \geq \max\{\theta, \theta^K\}$ when $\theta > 0$. Furthermore, given two matrices \mathbf{A} , \mathbf{B} such that $f_K(\sigma_1(\mathbf{B})) \leq f_K(\sigma_1(\mathbf{A}))$, we then have for any $\theta > 0$,

$$f_K(\sigma_1(\theta \mathbf{B})) = f_K(\theta \sigma_1(\mathbf{B})) \leq g_0(\theta) \cdot f_K(\sigma_1(\mathbf{B})) \leq g_0(\theta) \cdot \sigma_1(\mathbf{A}).$$

By equation 3.4, we finally arrive at

$$\sum_k^K \mathbf{D}[\theta; \mathbf{B}_k] \leq \mathbf{D}\left[g_0(\theta); \sum_k^K \mathbf{A}_k\right] + (K-1) \cdot \mathbf{\Lambda}\left\{0, 0, \log \frac{1}{2!}, \log \frac{1}{3!}, \dots\right\},$$

where \mathbf{A}_k ($1 \leq k \leq K$) are the fixed matrices such that $f_K(\sigma_1(\mathbf{B}_k)) \leq \sigma_1(\mathbf{A}_k)$, respectively.

Acknowledgments

We thank the anonymous reviewers and the editors for their valuable comments and suggestions. This work is partially supported by the Fundamental Research Funds for the Central Universities (DUT13RC(3)068 and DUT16LK05), the National Natural Science Foundation of China (11501079, 11401076, and 61473328), and Australian Research Council Projects (FT-130101457 and DP-140102164).

References

- Ahlsvede, R., & Winter, A. (2002). Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory*, 48(3), 569–579.
- Bian, W., & Tao, D. (2014). Asymptotic generalization bound of Fisher's linear discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(12), 2325–2337.
- Cantrell, R., & Cosner, C. (2004). *Spatial ecology via reaction-diffusion equations*. Hoboken, NJ: Wiley.
- Dai, G., Ma, R., Wang, H., Wang, F., & Xu, K. (2015). Partial differential equations with Robin boundary condition in online social networks. *Discrete and Continuous Dynamical Systems-Series B*, 20(6), 1609–1624.
- Hein, M., & Bühler, T. (2010). An inverse power method for nonlinear eigenproblems with applications in 1-spectral clustering and sparse PCA. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, & A. Culotta (Eds.), *Advances in neural information processing systems*, 24 (pp. 847–855). Red Hook, NY: Curran.

- Hein, M., & Setzer, S. (2011). Beyond spectral clustering: Tight relaxations of balanced graph cuts. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems*, 25 (pp. 2366–2374). Red Hook, NY: Curran.
- Higham, N. (2008). *Functions of matrices: Theory and computation*. Philadelphia: SIAM.
- Hsu, D., Kakade, S., & Zhang, T. (2012). Tail inequalities for sums of random matrices that depend on the intrinsic dimension. *Electronic Communications in Probability*, 17(14), 1–13.
- Jost, L., Setzer, S., & Hein, M. (2014). Nonlinear eigenproblems in data analysis: Balanced graph cuts and the ratioidca-prox. In S. Dahlke, W. Dahmen, M. Griebel, W. Hackbusch, K. Ritter, R. Schneider . . . H. Yserentant (Eds.), *Extraction of quantifiable information from complex systems* (pp. 263–279). New York: Springer.
- Mackey, L., Jordan, M., Chen, R., Farrell, B., & Tropp, J. (2014). Matrix concentration inequalities via the method of exchangeable pairs. *Annals of Probability*, 42(3), 906–945.
- Marshall, A., Olkin, I., & Arnold, B. (2010). *Inequalities: Theory of majorization and its applications*. New York: Springer Science & Business Media.
- Muir, D., & Mscis-Flogel, T. (2015). Eigenspectrum bounds for semirandom matrices with modular and spatial structure for neural networks. *Physical Review E*, 91(4), 042808.
- Rajan, K., & Abbott, L. (2006). Eigenvalue spectra of random matrices for neural networks. *Physical Review Letters*, 97(18), 188104.
- Tropp, J. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4), 389–434.
- Vershynin, R. (2010). *Introduction to the non-asymptotic analysis of random matrices*. arXiv preprint arXiv:1011.3027

Received May 30, 2016; accepted July 25, 2016.