Matrix Infinitely Divisible Series: Tail Inequalities and Applications in Optimization

Chao Zhang *Member, IEEE,* Xianjie Gao, Min-Hsiu Hsieh* *Senior Member, IEEE,* Hanyuan Hang, Dacheng Tao, *Fellow, IEEE*

***Abstract*—In this paper, we study tail inequalities of the largest eigenvalue of a matrix infinitely divisible (i.d.) series, which is a finite sum of fixed matrices weighted by i.d. random variables. We obtain several types of tail inequalities, including Bennett-type and Bernstein-type inequalities. This allows us to further bound the expectation of the spectral norm of a matrix i.d. series. Moreover, by developing a new lower-bound function for $Q(s) = (s+1)\log(s+1) - s$ that appears in the Bennett-type inequality, we derive a tighter tail inequality of the largest eigenvalue of the matrix i.d. series than the Bernstein-type inequality when the matrix dimension is high. The resulting lower-bound function is of independent interest and can improve any Bennett-type concentration inequality that involves the function $Q(s)$. The class of i.d. probability distributions is large and includes Gaussian and Poisson distributions, among many others. Therefore, our results encompass the existing work [1] on matrix Gaussian series as a special case. Lastly, we show that the tail inequalities of a matrix i.d. series have applications in several optimization problems including the chance constrained optimization problem and the quadratic optimization problem with orthogonality constraints.**

***Index Terms*—Random matrix, tail inequality, infinitely divisible distribution, largest eigenvalue, optimization**

I. Introduction

Random matrices have been widely used in many machine learning and information theory problems, *e.g.,* compressed sensing [2, 3, 4], coding theory [5], kernel method [6], estimation of covariance matrices [7, 8], and quantum information theory [9, 10, 11]. In particular, sums of random matrices and the tail behavior of their extreme eigenvalues (or singular values) are of significant interest in theoretical studies and practical applications (*cf.* [12]). Ahlswede and Winter presented a large-deviation inequality for the extreme eigenvalues of sums of random matrices [13]. Tropp improved upon their results using Lieb's concavity theorem [1]. Hsu *et al.* provided tail inequalities for sums of random matrices that depend on intrinsic dimensions instead of explicit matrix dimensions [14]. By introducing the concept of effective rank, Minsker extended Bernstein's concentration inequality for random matrices [15] and refined the results in [14]. There have also been many other works on the eigenproblems of random matrices (*cf.* [16, 17, 18, 19, 20]), and the list provided here is incomplete.

A simple form of sums of random matrices can be expressed as $\sum_k \xi_k \mathbf{A}_k$ with random variables ξ_k and fixed matrices $\mathbf{A}_k$. This form has played an important role in recent works on neural networks [21], kernel methods [22] and deep learning [23], where the original weighted (or projection) matrices can be replaced with structured random matrices, such as circulant and Toeplitz matrices with Gaussian or Bernoulli entries. Note that these two distributions, along with uniform distributions and Rademacher distributions, belong to the family of sub-Gaussian distributions[1], and many techniques dedicated to sub-Gaussian random matrices have been developed (*e.g.,* [1, 14]). However, to the best of our knowledge, random matrix research beyond that is still very limited.

The tail behavior of $\|\sum_k \xi_k \mathbf{A}_k\|$, where $\|\mathbf{A}\|$ stands for the spectral norm of the matrix $\mathbf{A}$, is strongly related to several optimization problems, including the Procrustes problem and the quadratic assignment problem (*cf.* [24, 25]). Nemirovski analyzed efficiently computable solutions to these optimization problems [24], and showed that the tail behavior of $\|\sum_k \xi_k \mathbf{A}_k\|$ provides answers to 1) the safe tractable approximation of chance constrained linear matrix inequalities, and 2) the quality of semidefinite relaxations of a general quadratic optimization problem. He also proved a tail bound for $\|\sum_k \xi_k \mathbf{A}_k\|$, where $\{\xi_k\}$ obey either distributions supported on $[-1, 1]$ or Gaussian distributions with *unit* variance, and presented a conjecture for the "optimal" expression of the tail bound [24]. Anthony So applied the non-commutative Khintchine's inequality to achieve a solution to Nemirovski's conjecture [25]. Note that the aforementioned results assume that $\{\xi_k\}$ obey distributions supported on $[-1, 1]$ or Gaussian distributions with *unit* variance. These assumptions will not always be satisfied in practice, and it is advantageous to explore whether these efficiently computable optimization solutions would also hold in a broader setting. We answer this question in the affirmative in this paper.

In this work, we study and prove tail bounds for the random matrix $\sum_k \xi_k \mathbf{A}_k$, where random variables $\{\xi_k\}$ are infinite

C. Zhang and X. Gao are with the School of Mathematical Sciences, Dalian University of Technology, Dalian, Liaoning, 116024, P.R. China. E-mail: chao.zhang@dlut.edu.cn, xianjiegao@foxmail.com.

M.-H. Hsieh is with Centre for Quantum Software and Information, University of Technology Sydney, Sydney NSW 2007, Australia. E-mail: Min-Hsiu.Hsieh@uts.edu.au.

H. Hang is with Institute of Statistics and Big Data, Renmin University of China, Beijing, 100872, P.R. China. E-mail: hans2017@ruc.edu.cn.

D. Tao is with the School of Information Technologies, The University of Sydney, Darlington, NSW 2008, Australia. E-mail: dacheng.tao@sydney.edu.au.

This work is partially supported by the Fundamental Research Funds for the Central Universities: DUT17LK46 and the National Natural Science Foundation of China: 11401076 and 61473328. MH is supported by an ARC Future Fellowship under Grant FT140100574.

*Corresponding author

[1] A random variable ξ is said to be sub-Gaussian if its moment generating function (mgf) satisfies $\mathbb{E}[e^{\theta\xi}] \leq e^{\theta^2 c^2}$ $(\forall \theta \in \mathbb{R})$, where c is an absolute constant.

divisible distributions. The class of infinitely divisible (i.d.) distributions includes Gaussian distributions, Poisson distributions, stable distributions and compound Poisson distributions as special cases (*cf.* [26, 27]). In recent years, techniques developed for i.d. distributions have been employed in important applications in the fields of image processing [28] and kernel methods [29]. Note that there is no intersection between sub-Gaussian distributions and i.d. distributions except for Gaussian distributions (*cf.* Lemma 5.5 of [19]). We therefore believe that our works on random matrix with respect to i.d. distributions will complement earlier results for sub-Gaussian distributions and provide useful applications in the fields of learning and optimization, and beyond.

A. Overview of the Main Results

There are three main contributions of this paper: 1) we obtain tail inequalities for the largest eigenvalue of the matrix infinitely divisible (i.d.) series $\sum_k \xi_k \mathbf{A}_k$, where the ξ_k are i.d. random variables; 2) we construct a piecewise function to bound the function $Q(s) = (s+1)\log(s+1) - s$ from below when $s \in (0, c]$ for any given $1 < c < +\infty$, and the new lower bound function is the tightest up to date; and 3) we show that the tail inequalities of matrix i.d. series provide efficiently computable solutions to several optimization problems.

First, we develop a matrix moment-generating function (mgf) bound for i.d. distributions as the starting point for deriving the subsequent tail inequalities for the matrix i.d. series. Then, we derive the tail inequality given in (5) for the matrix i.d. series, which is difficult to compute because of the integral of an inverse function. Therefore, by introducing the additional condition that the Lévy measure has a bounded support, we simplify the aforementioned result into a Bennett-type tail inequality [*cf.* (6)] that contains the function $Q(s) = (s+1)\log(s+1) - s$, and we also replace $Q(s)$ with $B(s) = \frac{s^2}{2(1+s/3)}$ to obtain a Bernstein-type tail inequality [*cf.* (10)] for the matrix i.d. series. In addition, we bound the expectation of the spectral norm of the matrix i.d. series.

Since $B(s)$ cannot bound $Q(s)$ from below sufficiently tightly when s is large (*cf.* Fig. 1), we introduce another function $H_P(s)$ [*cf.* (16)] to bound $Q(s)$ from below more tightly than $B(s)$ when $s \in (0.8831, c]$ for any $1 < c < +\infty$ (*cf.* Remark 3.3). Although $H_P(s)$ is a piecewise function, all sub-functions of $H_P(s)$ share the simple form $\beta_0 s^{\tau_n}$ (where $\beta_0 = 2\log 2 - 1$) and thus have a low computational cost, and the subdomains of $H_P(s)$ can be arbitrarily selected as long as points 1 and c are included in the ordered sequence P as the smallest and largest elements, respectively. Based on $H_P(s)$ (especially with $P = \{1, c\}$), we obtain another type of tail inequality for matrix i.d. series that is tighter than the Bernstein-type result given in (10) when $\frac{Rt}{\rho(\sigma^2 V)} > 0.8831$.[2] We show that the tail result based on $H_P(s)$ provides a tighter upper bound on the largest eigenvalue of a matrix i.d. series than is possible with the Bernstein-type result when the matrix dimension is high. The results regarding $Q(s)$ and $H_P(s)$ are applicable for any Bennett-type concentration inequality that involves the function $Q(s)$.

Using the resulting tail bounds for random i.d. series, we study the properties of two optimization problems: chance constrained optimization problems and quadratic optimization problems with orthogonality constraints, which covers several well-studied optimization problems as special cases, *e.g.,* the Procrustes problem and the quadratic assignment problem. Although these problems have been exhaustively explored in the works [24, 25], their results are built under the assumption that ξ_k obey either distributions supported on $[-1, 1]$ or Gaussian distributions with *unit* variance, which restricts the feasibility of the results in practical problems that do not satisfy the assumption. By using the tail inequalities for random i.d. series to resolve an extension of Nemirovski's conjecture (*cf.* Conjecture 4.1), we show that the results obtained in [24, 25] are also valid in the i.d. scenario, where ξ_k obey i.d. distributions instead of distributions supported on $[-1, 1]$ or Gaussian distributions.

The remainder of this paper is organized as follows. Section II introduces necessary preliminaries on i.d. distributions and Section III presents the main results of this paper. In Section IV, we study the application of random i.d. series in a number of optimization problems. Section V concludes the paper. In the appendix, we provide a detailed introduction to the Lévy measure (part A) and prove the main results of this paper (part B).

II. Preliminaries on Infinitely Divisible Distributions

In this section, we first introduce several definitions related to infinitely divisible (i.d.) distributions and then present the matrix mgf inequality for i.d. distributions.

A. Infinitely Divisible Distributions

A random variable ξ has an i.d. distribution if for any $n > 1$, there exists a sequence $\{\xi_n^{(1)}, \cdots, \xi_n^{(n)}\}$ of independent and identically distributed (i.i.d.) random variables such that ξ has the same distribution as $\xi_n^{(1)} + \cdots + \xi_n^{(n)}$. Equivalently, i.d. distributions can be defined by means of a characteristic exponent, as follows.

Definition 2.1: Let $\phi(\theta)$ be the characteristic exponent of a random variable ξ:

$$\phi(\theta) := \log \mathbb{E}\{\mathrm{e}^{i\theta\xi}\} = \log \int_{-\infty}^{+\infty} \mathrm{e}^{i\theta\xi} dP(\xi), \quad \theta \in \mathbb{R}.$$

The distribution of ξ is said to be i.d. if for any $n \in \mathbb{N}$, there exists a characteristic exponent $\phi_n(\theta)$ such that

$$\phi(\theta) = \underbrace{\phi_n(\theta) + \cdots + \phi_n(\theta)}_{n}.$$

Now, we need to introduce the concept of the Lévy measure.

Definition 2.2 (Lévy Measure): A Borel measure ν defined on $\mathbb{R}$ is said to be a Lévy measure if it satisfies

$$\int_{\mathbb{R}} \min\{u^2, 1\} \nu(du) < \infty \ \text{ and } \nu(\{0\}) = 0. \tag{1}$$

[2] In general, the tail inequality $\mathbb{P}\{\xi > t\}$ describes the probability characteristics of the event in which the value of a random variable ξ is greater than a given positive constant t. Consequently, the tail inequality provides more useful information in the case of $\frac{Rt}{\rho(\sigma^2 V)} > 0.8831$ than in the case of $\frac{Rt}{\rho(\sigma^2 V)} \leq 0.8831$.

The Lévy measure describes the expected number of jumps of a certain height in a time interval of *unit* length; a more detailed explanation is given in Appendix A. The following theorem provides a sufficient and necessary condition for i.d. distributions:

Theorem 2.1 (Lévy-Khintchine Theorem): A real-valued random variable ξ is i.d. if and only if there exists a triplet (b, σ^2, ν) such that for any $\theta \in \mathbb{R}$, the characteristic exponent $\phi(\theta)$ is of the form

$$\phi(\theta) = ib\theta - \frac{\sigma^2\theta^2}{2} + \int_{\mathbb{R}} \left(\mathrm{e}^{i\theta u} - 1 - i\theta u \mathbf{1}_{(|u|<1)}\right)\nu(du), \quad (2)$$

where $b \in \mathbb{R}$, $\sigma \geq 0$ and ν is a Lévy measure.
This theorem states that an i.d. distribution can be characterized by the triplet (b, σ^2, ν). Refer to [26, 27] for details.

B. Matrix Inequalities for Infinitely Divisible Distributions

Let the symbol $\preceq$ denote the semidefinite order on self-adjoint matrices. For any real functions f and g, the transfer rule states that if $f(a) \leq g(a)$ for any $a \in I$, then $f(\mathbf{A}) \preceq g(\mathbf{A})$ when the eigenvalues of the semidefinite matrix $\mathbf{A}$ lie in I. Below, we present the matrix mgf bound for i.d. distributions as the starting point for deriving the desired tail results for matrix i.d. series.

Lemma 2.1: Let ξ be an i.d. random variable with the triplet (b, σ^2, ν), and suppose that $\mathbb{E}\xi = 0$. Given a fixed self-adjoint matrix $\mathbf{A}$ with $\lambda_{\max}(\mathbf{A}) \leq 1$, it holds that for any $0 < \theta \leq M$,

$$\mathbb{E}\mathrm{e}^{\xi\theta\mathbf{A}} \preceq \mathrm{e}^{\Phi(\theta)\cdot\mathbf{A}^2}, \quad (3)$$

where $\lambda_{\max}(\cdot)$ stands for the largest eigenvalue, $M := \sup\{\theta \geq 0 : \mathbb{E}\,\mathrm{e}^{\theta|\xi|} < +\infty\}$ and

$$\Phi(\theta) := \frac{\sigma^2\theta^2}{2} + \int_{\mathbb{R}} \left(\mathrm{e}^{\theta|u|} - \theta|u| - 1\right)\nu(du). \quad (4)$$

The proof of this lemma is given in Appendix B-A. Note that if the Lévy measure ν is the *zero* measure, then the mgf result given in (3) is analogous to the mgf result $\mathbb{E}\mathrm{e}^{\xi\theta\mathbf{A}} = \mathrm{e}^{\theta^2\mathbf{A}^2/2}$ ($\forall\theta \in \mathbb{R}$) when ξ is Gaussian (*cf.* Lemma 4.3 of [1]).

III. Tail Inequalities for Matrix Infinitely Divisible Series

In this section, we first present two types of tail inequalities for matrix i.d. series: Bennett-type and Bernstein-type inequalities. By analyzing the characteristics of the function $Q(s) = (s+1)\cdot\log(s+1) - s$ that appears in the Bennett-type result, we introduce a piecewise function $H(s)$ to bound Q(s) from below and thus obtain a new tail inequality for matrix i.d. series. We also study the upper bound of the expectation of $\|\sum_k \xi_k\mathbf{A}_k\|$.

A. Tail Inequalities for Matrix Infinitely Divisible Series

By using the the matrix mgf bound (3), we first obtain the tail inequality for the matrix i.d. series $\sum_k \xi_k\mathbf{A}_k$:

Theorem 3.1: Let $\mathbf{A}_1, \cdots, \mathbf{A}_K$ be fixed d-dimensional self-adjoint matrices with $\lambda_{\max}(\mathbf{A}_k) \leq 1$ ($k = 1, \cdots, K$), and let $\xi_1, \cdots, \xi_K$ be independent centered i.d. random variables with the triplet (b, σ^2, ν) and $M := \sup\{\theta \in \mathbb{R} : \mathbb{E}\,\mathrm{e}^{\theta|\xi|} < +\infty\}$. Define $\rho := \lambda_{\max}\big(\sum_k \mathbf{A}_k^2\big)$. Then for all $0 < t < \frac{\alpha(M^-)}{\rho}$, we have

$$\begin{aligned} &2\mathbb{P}\left\{\lambda_{\max}\left(\sum_k \xi_k\mathbf{A}_k\right) > t\right\} = \mathbb{P}\left\{\left\|\sum_k \xi_k\mathbf{A}_k\right\| > t\right\} \\ &\leq 2d\exp\left(-\rho\cdot\int_0^{t/\rho}\alpha^{-1}(s)ds\right), \end{aligned} \quad (5)$$

where $\alpha(M^-)$ is the left limit at M, and $\alpha^{-1}(s)$ is the inverse of

$$\alpha(s) := \sigma^2 s + \int_{\mathbb{R}} |u|(\mathrm{e}^{s|u|} - 1)\nu(du), \quad 0 < s < M.$$

The proof of this theorem is given in Appendix B-B.

Remark 3.1: Since the matrices $\mathbf{A}_k$ ($1 \leq k \leq K$) are self-adjoint, the matrix $\sum_k \mathbf{A}_k^2$ is self-adjoint and positive semidefinite. Therefore, ρ is non-negative and the above result is non-trivial.

Considering the difficulties that arise in computing the function $\alpha(s)$ and its inverse $\alpha^{-1}(s)$, we introduce the additional condition that ν has a bounded support to simplify the above result, which leads to the following corollary.

Corollary 3.1: If ν has a bounded support with $R = \inf\{\alpha > 0 : \nu(\{u : |u| > \alpha\}) = 0\}$, then for any $t > 0$,

$$\begin{aligned} &2\mathbb{P}\left\{\lambda_{\max}\left(\sum_k \xi_k\mathbf{A}_k\right) > t\right\} = \mathbb{P}\left\{\left\|\sum_k \xi_k\mathbf{A}_k\right\| > t\right\} \\ &\leq 2d\cdot\exp\left(-\frac{\rho(\sigma^2+V)}{R^2}\cdot Q\left(\frac{Rt}{\rho(\sigma^2+V)}\right)\right), \end{aligned} \quad (6)$$

where $V := \int_{\mathbb{R}} |u|^2\nu(du)$, and

$$Q(s) := (1+s)\cdot\log(1+s) - s. \quad (7)$$

The proof of this corollary is given in Appendix B-C.

Roughly speaking, the condition that ν has a bounded support means that large jumps may not occur on the path of the Lévy process that is generated from the i.d. distribution with triplet (b, σ^2, ν). Refer to Appendix A for the explanation for this condition.

Note that the tail inequality (6) is similar in form to the matrix Bennett result (*cf.* Theorem 6.1 of [1]). Following the classical method of bounding $Q(s)$ from below, the Bernstein-type result can be derived based on the fact that

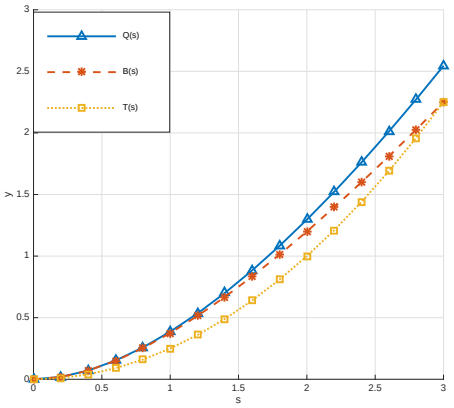
$$Q(s) \geq B(s) \geq T(s), \quad s \geq 0, \quad (8)$$

where

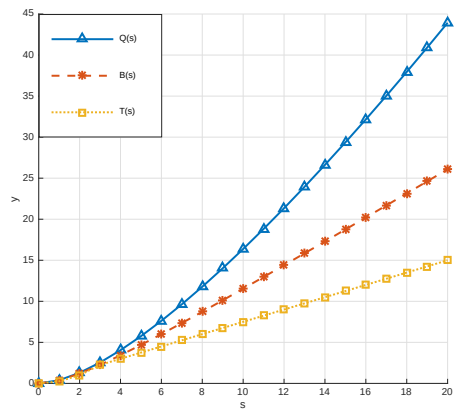
$$B(s) := \frac{s^2}{2(1+s/3)}; \;\; T(s) := \begin{cases} 3s/4, & s \geq 3; \\ s^2/4, & 0 < s < 3. \end{cases} \quad (9)$$

As shown in Fig. 1, the function $B(s) = \frac{s^2}{2(1+s/3)}$ can tightly bound $Q(s)$ from below when s is close to the *origin*, whereas there will be a large discrepancy between $Q(s)$ and $B(s)$ when s is far from the *origin*. This is because $B(s)$ is derived from the Taylor expansion at the point $s = 0$ (*cf.* Chapter 2.7 of [30]). To facilitate the analysis of $Q(s)$, the function $B(s)$ is relaxed to a looser lower-bound function

$T(s)$, which is a piecewise function with the following sub-functions: $s^2/4$ when $s \in (0,3)$; and $3s/4$ when $s \in [3,\infty)$. Although the function $T(s)$ does not bound $Q(s)$ sufficiently tightly, the result presented in (15) below shows that $T(s)$ provides the same rate of growth as $Q(s)$ when s is close to the *origin* or approaches *infinity*. This phenomenon suggests that the coefficients $3/4$ and $1/4$ of the sub-functions $3s/4$ and $s^2/4$, respectively, are probably not sufficiently well-tuned.



(a) $s \in (0,3]$



(b) $s \in (0,20]$

Fig. 1. The function curves of $Q(s)$, $B(s)$ and $T(s)$.

Corollary 3.2: Let $\xi_1, \cdots, \xi_K$ be independent i.d. random variables satisfying the conditions in Corollary 3.1. Then for any $t > 0$,

$$\begin{aligned} &2\mathbb{P}\left\{\lambda_{\max}\left(\sum_k \xi_k \mathbf{A}_k\right) > t\right\} = \mathbb{P}\left\{\left\|\sum_k \xi_k \mathbf{A}_k\right\| > t\right\} \\ \le& 2d \cdot \exp\left(-\frac{3}{2}\cdot\frac{t^2}{3\rho(\sigma^2+V)+Rt}\right) \qquad (10) \\ \le& \begin{cases} 2d\cdot\exp\left(-\frac{3}{4}\cdot\frac{t}{R}\right), & \text{if } Rt > 3\rho(\sigma^2+V); \\ 2d\cdot\exp\left(-\frac{t^2}{4\rho(\sigma^2+V)}\right), & \text{if } 0 < Rt \le 3\rho(\sigma^2+V). \end{cases} \end{aligned}$$

This corollary shows that the probability of the event $\|\sum_k \xi_k \mathbf{A}_k\| > t$ is bounded by $O(\mathrm{e}^{-c_1 t})$ when t is large and that its upper bound is of the form $O(\mathrm{e}^{-c_2 t^2})$ when t is small.

Recalling Inequality (4.9) of [1], the expectation $\mathbb{E}\|\sum_k \xi_k \mathbf{A}_k\|$ for a random Gaussian series is bounded by the term $O[\sqrt{\log(c\cdot d)}]$. In a similar way, we use the tail bound presented in (10) to obtain an upper bound on $\mathbb{E}\|\sum_k \xi_k \mathbf{A}_k\|$ for a random i.d. series.

Theorem 3.2: Let $\xi_1, \cdots, \xi_K$ be independent i.d. random variables satisfying the conditions in Corollary 3.1. Then

$$\mathbb{E}\left\|\sum_k \xi_k \mathbf{A}_k\right\| \le \frac{3R}{4}\cdot\log\left(2d\cdot \mathrm{e}^{1+\frac{9\rho^2(\sigma^2+V)^2}{2R^2}}\right). \qquad (11)$$

Because of the existence of the Lévy measure ν, the upper bound on $\mathbb{E}\|\sum_k \xi_k \mathbf{A}_k\|$ for a random i.d. series is of the form $O[\log(c\cdot d)]$, which differs from the Gaussian bound of $O[\sqrt{\log(c\cdot d)}]$. Recalling the Lévy-Itô decomposition (*cf.* [27]), the higher expectation bound for a matrix i.d. series arises from the existence of the compound Poisson (with drift) components of the i.d. distribution.

Remark 3.2:

Note that the aforementioned tail results for matrix i.d. series can be generalized to the scenario of sums of independent i.d. random matrices $\mathbf{X}_1, \cdots, \mathbf{X}_K$, all of whose entries are i.d. random variables with the generating triplet (b, σ^2, ν). As a starting point, we first obtain the mgf bound for the self-adjoint i.d. random matrix $\mathbf{X}$ with (b, σ^2, ν) and $\lambda_{\max}(\mathbf{X}) \le 1$:

$$\mathbb{E}\mathrm{e}^{\theta\mathbf{X}} \preceq \mathrm{e}^{\Phi(\theta)\cdot\mathbb{E}(\mathbf{X}^2)}, \quad \forall\, 0 < \theta \le M, \qquad (12)$$

which can be proven in a manner similar to Lemma 2.1. We then arrive at upper bounds on $\mathbb{P}\{\|\sum_k \mathbf{X}_k\| > t\}$ and $\mathbb{E}\|\sum_k \mathbf{X}_k\|$ with the same forms as those of the proposed results for matrix i.d. series except that the term $\rho = \lambda_{\max}(\sum_k \mathbf{A}_k^2)$ is replaced by $\rho_0 = \lambda_{\max}(\sum_k \mathbb{E}(\mathbf{X}_k^2))$ [*cf.* (5), (6), (10), (17) and (11)]. These results can also be regarded as an extension of the existing vector-version results (*cf.* [31, 32]).

B. A Lower-Bound Function of $Q(s)$

As discussed above, both $B(s)$ and $T(s)$ are lower bound functions for $Q(s)$, but they do not bound $Q(s)$ sufficiently tightly when s is far from the *origin* (*cf.* Fig. 1) because they stem from the Taylor expansion at the *origin*. We adopt a more direct strategy to analyze the behavior of the function $Q(s)$; for earlier discussions on this topic, refer to [33, 34].

We consider the following inequality:

$$(s+1)\cdot\log(s+1) - s \ge \beta\cdot s^{\tau}, \quad \forall\, s > 0, \qquad (13)$$

where the parameter β is expected to be a constant independent of s such that $\beta\cdot s^{\tau}$ bounds $Q(s)$ from below as tightly as possible. For any $s \in (0,1)\cup(1,+\infty)$, define

$$\tau(\beta, s) := \frac{\log\big((s+1)\log(s+1) - s\big) - \log(\beta)}{\log(s)}. \qquad (14)$$

Then, it follows L'Hospital's rule that

$$\lim_{s\to 0^+}\tau(\beta,s) = 2 \ \text{ and } \ \lim_{s\to+\infty}\tau(\beta,s) = 1, \ \ \forall\beta > 0. \qquad (15)$$

The two limits in (15) suggest that piecewise function $T(s)$ indeed captures the rate of growth of the function $Q(s)$ as s approaches either the *origin* or *infinity*.

Now, we must choose the parameter β. As shown in Fig. 2, the function $\tau(\beta, s)$ is sensitive to the choice of β, and the value of $\tau(\beta, s)$ will vary dramatically near the point $s = 1$ if parameter β is not chosen well. Therefore, we should select a β such that the variation of $\tau(\beta, s)$ near $s = 1$ is kept as small as possible, *i.e.,* such that the discrepancy between $\tau(\beta, 1^-)$ and $\tau(\beta, 1^+)$ is minimized. The follow lemma is also derived from L'Hospital's rule:

Lemma 3.1: Let $\beta_0 = 2\log 2 - 1$. Then,

$$\lim_{s\to 1^-} \tau(\beta_0, s) = \lim_{s\to 1^+} \tau(\beta_0, s) = \frac{\log 2}{2\log 2 - 1}.$$

This lemma shows that with the parameter choice $\beta = 2\log 2 - 1$, the point $s = 1$ is a removable discontinuity of the function $\tau(\beta, s)$; *i.e.,* $\tau(\beta, 1^-) = \tau(\beta, 1^+)$. In other words, if we add a supplementary definition of $\tau(\beta, 1) := \frac{\log 2}{2\log 2 - 1}$, the resulting function $\tau(\beta, s)$ will be continuous on the domain $(0, +\infty)$. Therefore, parameter β should be selected such that $\beta = \beta_0 = 2\log 2 - 1$.

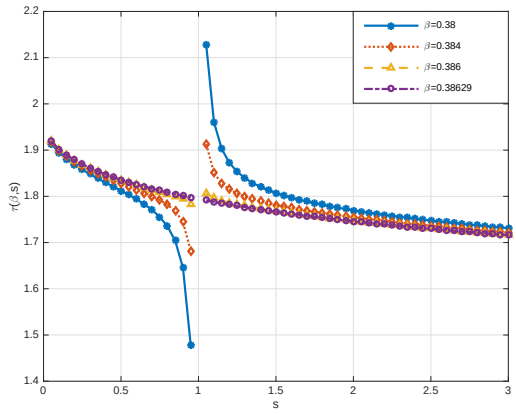


Fig. 2. The function curves of $\tau(\beta, s)$ w.r.t. different β settings

By using the function $\tau(\beta_0, s)$, we can develop another lower-bound function for $Q(s)$ as follows.

Proposition 3.1: Given an arbitrary positive constant $c > 1$ and an integer $N > 1$, let $P = \{p_0, p_1, \cdots, p_N\}$ be an ordered sequence such that $1 = p_0 < p_1 < \cdots < p_{N-1} < p_N = c$, and define

$$H_P(s) := \begin{cases} \beta_0 \cdot s^2, & 0 < s \le p_0; \\ \beta_0 \cdot s^{\tau_1}, & p_0 < s \le p_1; \\ \beta_0 \cdot s^{\tau_2}, & p_1 < s \le p_2; \\ \quad\vdots & \quad\vdots \\ \beta_0 \cdot s^{\tau_n}, & p_{N-1} < s \le p_N, \end{cases} \tag{16}$$

where $\beta_0 = 2\log 2 - 1$ and $\tau_n := \tau(\beta_0, p_n)$ $(n = 1, 2, \cdots, N)$. Then, for all $s \in (0, c]$, we have $Q(s) \ge H_P(s) \ge H_{\{1,c\}}(s)$, where the first equality holds when $s = p_0$ or $s = p_N$; and the second equality holds when $P = \{1, c\}$.

As suggested by this result, a piecewise function $H_P(s)$ to bound $Q(s)$ from below can be built when s has a bounded domain $(0, c]$ by means of the following steps: (i)

1) Let $\beta_0 = 2\ln 2 - 1$, and select a constant c to form an interval $(0, c]$.
2) Select an integer $N > 1$ and an ordered sequence $P := \{p_0, p_1, \cdots, p_N\}$ such that $1 = p_0 < p_1 < \cdots < p_N = c$.
3) If $s \in (0, 1]$, then $H_P(s) = \beta_0 s^2$; if $s \in (p_{n-1}, p_n]$, then $H_P(s) = \beta_0 s^{\tau_n}$, where $\tau_n = \tau(\beta_0, p_n)$ $(n = 1, 2, \cdots, N)$.

The resulting function $H_P(s)$ has the following characteristics:

- There is no additional restriction on the choice of the constant c, the integer N and the points $p_1, p_2, \cdots, p_{N-1}$ other than $1 = p_0 < p_1 < \cdots < p_N = c$. This means that suitable parameters c, N and $\{p_1, p_2, \cdots, p_{N-1}\}$ can be chosen in accordance with the requirements of various practical problems.
- Although $H_P(s)$ is a piecewise function, all parts of $H_P(s)$ share the same coefficient $\beta_0 = 2\log 2 - 1$, and the parameters τ_n are the values of function $\tau(\beta_0, s)$ at the partition points p_n $(n = 1, 2, \cdots, N)$. Therefore, the computation of $H_P(s)$ has a low cost.
- For any choice of P, the piecewise function $H_P(s)$ has the same form $\beta_0 \cdot s^2$ when $s \in (0, 1]$. In particular, $H_{\{1,c\}}(s)$ (*i.e.*, with $P = \{1, c\}$) is a continuous function on $(0, c)$, and the difference between $H_{\{1,c\}}(s)$ and $H_P(s)$ is not significant for any other choice of P (*cf.* Fig. 3). Hence, $H_c(s) := H_{\{1,c\}}(s)$ can be adopted as the lower-bound function for $Q(s)$ if there are no additional requirements on the ordered sequence P.

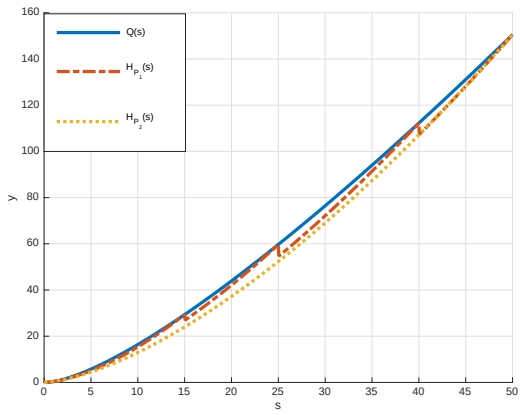
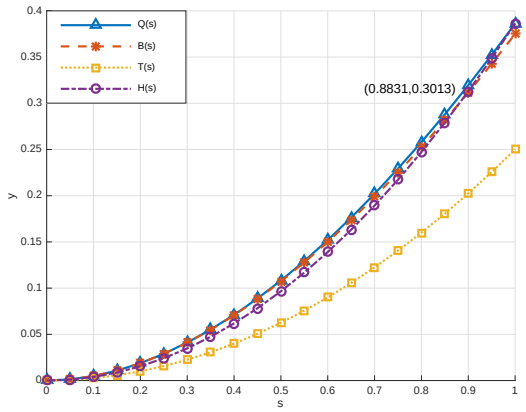


Fig. 3. The function curves of H_P w.r.t. different P settings, where $P_1 = \{1, 15, 25, 40, 50\}$ and $P_2 = \{1, 50\}$. Although the function H_{P_1} is closer to $Q(s)$ than H_{P_2} is, the curve of H_{P_1} is not continuous and the discrepancy between H_{P_1} and H_{P_2} is not significant.

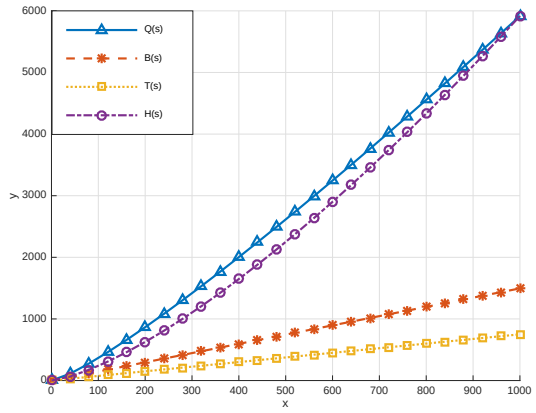
Remark 3.3:

As shown in Fig. 4, the lower-bound function $H_c(s)$ performs better than the function $B(s)$, which is derived from the Taylor expansion, when $s \in (0.8831, c]$; moreover, although $B(s)$ bounds $Q(s)$ more tightly than $H_c(s)$ does when $s \in (0, 0.8831]$, there is only a slight discrepancy between $H_c(s)$ and $B(s)$ on this interval.[3] As a result, the

[3]The range of $s \in (0.8831, c)$ is the numerical solution to the inequality $H_c(s) > B(s)$.



(a) $s \in (0, 1]$



(b) $s \in (0, 1000]$

Fig. 4. The function curves of $Q(s)$, $B(s)$, $T(s)$ and $H_c(s)$ with $c = 1000$. The curves of $H_c(s)$ and $B(s)$ intersect approximately at the point $(0.8831, 0.3013)$, and the function $H_c(s)$ is closer to $Q(s)$ than $B(s)$ is when $0.8831 < s < 1000$.

method of bounding $Q(s)$ that is proposed in (13) is not only effective but also corrects for the shortcoming of the Taylor-expansion-based method (8), *i.e.,* the local approximation at the *origin*.

By recalling the tail inequality (6) and replacing the function $Q(s)$ with $H_c(s)$, we obtain, for any $0 < \frac{Rt}{\rho(\sigma^2+V)} \leq c$,

$$2\mathbb{P}\Big\{\lambda_{\max}\Big(\sum_k \xi_k \mathbf{A}_k\Big) > t\Big\} = \mathbb{P}\Big\{\Big\|\sum_k \xi_k \mathbf{A}_k\Big\| > t\Big\} \quad (17)$$

$$\leq \begin{cases} 2d \cdot \exp\left(-\frac{\beta_0}{\rho(\sigma^2+V)} \cdot t^2\right), & \text{if } 0 < \frac{Rt}{\rho(\sigma^2+V)} \leq 1; \\ 2d \cdot \exp\left(-\frac{\beta_0 \cdot R^{\tau_c-2}}{[\rho\cdot(\sigma^2+V)]^{\tau_c-1}} \cdot t^{\tau_c}\right), & \text{if } 1 < \frac{Rt}{\rho(\sigma^2+V)} \leq c, \end{cases}$$

where $\tau_c = \tau(\beta_0, c)$. As shown in Fig. 5, the above result provides a bound that is tighter than the one achieved by the Bernstein-type results in (10) when $\frac{Rt}{\rho(\sigma^2+V)} \in (0.8831, c)$, and is only slightly looser than the Bernstein-type bound based on $B(s)$ when $\frac{Rt}{\rho(\sigma^2+V)} \in (0, 0.8831)$.

Remark 3.4:

Since the function $H_c(s)$ is defined on the bounded interval $(0, c]$, the result given in (17) cannot be used to analyze the asymptotic behavior of $\mathbb{P}\{\lambda_{\max}(\sum_k \xi_k \mathbf{A}_k) > t\}$ as t goes to *infinity*. However, since $H_c(s)$ bounds $Q(s)$ from below more tightly than $B(s)$ (or $T(s)$) does on the bounded domain $s \in (08831, c]$, the result given in (17) provides a more accurate description of the non-asymptotic behavior of $\mathbb{P}\{\lambda_{\max}(\sum_k \xi_k \mathbf{A}_k) > t\}$ when $\frac{Rt}{\rho(\sigma^2+V)} > 3$. The following alternative expressions for the Bernstein-type result given in (10) and the H_c-based result given in (17) can respectively be obtained: with probability at least $1 - \delta$,

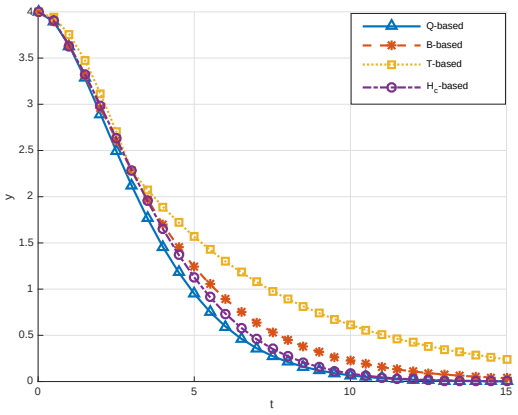


Fig. 5. The curves of Q-based, B-based, T-based and H_c-based tail bounds, where, for simplicity, the parameters are set as $d = 2$, $R = 4$ and $\rho(\sigma^2 + V) = 4$.

$$\lambda_{\max}\left(\sum_k \xi_k \mathbf{A}_k\right) \leq \frac{4R(\log 2d - \log \delta)}{3} \quad (18)$$

and

$$\lambda_{\max}\left(\sum_k \xi_k \mathbf{A}_k\right) \leq \left(\frac{\big(\log 2d - \log \delta\big)\big[\rho(\sigma^2 + V)\big]^{\tau_c - 1}}{\beta_0 R^{\tau_c - 2}}\right)^{\frac{1}{\tau_c}}.$$

These expressions suggest that $\lambda_{\max}\big(\sum_k \xi_k \mathbf{A}_k\big)$ is bounded by the term $O((\log d)^{\frac{1}{\tau_c}})$ with $1 < \tau_c < 2$, which is a tighter bound than the right-hand side of the Bernstein-type result (18) when the matrix dimension d is high.

IV. Applications in Optimization

In this section, we will show that the derived tail inequalities for random i.d. series can be used to solve two types of optimization problems: chance constrained optimization problems and quadratic optimization problems with orthogonality constraints. These optimization problems are reviewed in Section IV-A, and Nemirovski's conjecture [24] for efficiently computable solutions to these two optimization problems is introduced. We argue that the requirement in Nemirovski's conjecture is not practical, generalize the requirement using matrix i.d. series, and provide a solution to the extended version of Nemirovski's conjecture in Section IV-B. Lastly, we re-derive efficiently computable solutions to both types of optimization problems with a matrix i.d. series requirement in Section IV-C.

A. Relevant Optimization Problems

It has been pointed out in the pioneering work of [24] that the behavior of $\sum_k \xi_k \mathbf{A}_k$ is strongly related to the efficiently computable solutions to many optimization problems, *e.g.,* the chance constrained optimization problem and the quadratic optimization problem with orthogonality constraints. Several well-studied optimization problems are included in the latter as special cases, such as the Procrustes problem and the quadratic assignment problem. We begin with a brief introduction of these optimization problems.

1) Chance Constrained Optimization Problem: Consider the following chance constrained optimization problem (*cf.* [25]): given an N-dimensional vector $\mathbf{c} \in \mathbb{R}^N$ and an $\epsilon \in (0,1)$, find

$$\min_{\mathbf{x}\in\mathbb{R}^N} \mathbf{c}^T\mathbf{x} \quad \text{subject to} \tag{19}$$
$$\begin{cases} \mathbf{F}(\mathbf{x}) \preceq \mathbf{0}, & (a); \\ \mathbb{P}\Big\{\mathcal{A}_0(\mathbf{x}) - \sum_{k=1}^K \xi_k \mathcal{A}_k(\mathbf{x}) \succeq \mathbf{0}\Big\} \geq 1-\epsilon, & (b), \end{cases}$$

where $\mathbf{F} : \mathbb{R}^N \to \mathbb{R}^L$ is an efficiently computable vector-valued function with convex components; $\mathcal{A}_0, \mathcal{A}_1, \cdots, \mathcal{A}_K : \mathbb{R}^N \to \mathbb{S}^M$ are affine functions taking values in the space $\mathbb{S}^M$ of symmetric $M \times M$ matrices with $\mathcal{A}_0(\mathbf{x}) \succeq \mathbf{0}$ for all $\mathbf{x} \in \mathbb{R}^N$; and $\xi_1, \cdots, \xi_K$ are independent random variables with *zero* mean. The main challenge in solving this optimization lies in the chance constraint (19-b).

By letting $\mathcal{A}'_k(\mathbf{x}) = (\mathcal{A}_0(\mathbf{x}))^{-1/2}\mathcal{A}_k(\mathbf{x})(\mathcal{A}_0(\mathbf{x}))^{-1/2}$, we have

$$\mathbb{P}\Big\{\mathcal{A}_0(\mathbf{x}) - \sum_{k=1}^K \xi_k \mathcal{A}_k(\mathbf{x}) \succeq \mathbf{0}\Big\} = \mathbb{P}\Big\{\sum_{k=1}^K \xi_k \mathcal{A}'_k(\mathbf{x}) \preceq \mathbf{I}\Big\}.$$

It is subsequently necessary to find a sufficient condition for the inequality

$$\mathbb{P}\left\{\sum_{k=1}^K \xi_k \mathcal{A}'_k(\mathbf{x}) \preceq \mathbf{I}\right\} \geq 1-\epsilon, \tag{20}$$

and to guarantee that the condition can be efficiently computable in optimization. For example, So proposed the following condition [25]:

$$\sum_{k=1}^K (\mathcal{A}'_k(\mathbf{x}))^2 \preceq \gamma \mathbf{I} \quad \text{with} \quad \gamma = \gamma(\epsilon) > 0. \tag{21}$$

By using the Schur complement, it can be equivalently expressed as a linear matrix inequality:

$$\begin{bmatrix} \gamma\mathcal{A}_0(\mathbf{x}) & \mathcal{A}_1(\mathbf{x}) & \cdots & \mathcal{A}_K(\mathbf{x}) \\ \mathcal{A}_1(\mathbf{x}) & \gamma\mathcal{A}_0(\mathbf{x}) & & \\ \vdots & & \ddots & \\ \mathcal{A}_K(\mathbf{x}) & & & \gamma\mathcal{A}_0(\mathbf{x}) \end{bmatrix} \succeq \mathbf{0}. \tag{22}$$

If the constraint (19-b) is replaced with the inequality (22), the chance-constrained optimization problem will become tractable. To guarantee the validity of this replacement, it is necessary to consider the following problem:

(P1) Is the condition (21) sufficient for the inequality (20)?

2) Quadratic Optimization Problems with Orthogonality Constrains: Let $\mathbb{M}^{M\times N}$ be the space of $M \times N$ real matrices equipped with the trace inner product $\mathbf{X} \bullet \mathbf{Y} = \mathrm{tr}(\mathbf{X}^T\mathbf{Y}) = \mathrm{tr}(\mathbf{Y}^T\mathbf{X})$. Consider the following quadratic optimization problem:

$$\min_{\mathbf{X}\in\mathbb{M}^{M\times N}} \mathbf{X} \bullet \mathcal{A}\mathbf{X} \ \text{ subject to}$$
$$\begin{cases} \mathbf{X} \bullet \mathcal{B}_i\mathbf{X} \leq 1, \ \forall\, i = 1, \cdots, I; & (a) \\ \mathcal{C}\mathbf{X} = \mathbf{0}; & (b) \\ \|\mathbf{X}\| \leq 1, & (c) \end{cases} \tag{23}$$

where $\mathcal{A}, \mathcal{B}_1, \cdots, \mathcal{B}_I : \mathbb{M}^{M\times N} \to \mathbb{M}^{M\times N}$ are self-adjoint linear mappings (note that they can be represented as symmetric $MN \times MN$ matrices); $\mathcal{B}_1, \cdots, \mathcal{B}_I$ are positive semidefinite; $\mathcal{C} : \mathbb{M}^{M\times N} \to \mathbb{R}^L$ is a linear mapping (which can be represented as symmetric $L \times MN$ matrices); and $\|\mathbf{X}\|$ is the spectral norm of $\mathbf{X}$. As addressed in [24], this optimization problem covers many well-studied optimization problems with the orthogonality constraint $\mathbf{X}^T\mathbf{X} = \mathbf{I}$ as special cases, *e.g.,* the Procrustes problem and the quadratic assignment problem. By exploiting the structure of these problems, the orthogonality constraint $\mathbf{X}^T\mathbf{X} = \mathbf{I}$ can be relaxed to the constraint (23-c) without loss of generality.

The optimization problem can be directly tackled by using the semidefinite programming (SDP) relaxation:

$$\min \mathbf{D} \bullet \mathbf{Y} \ \text{ subject to}$$
$$\begin{cases} \mathbf{B}_i \bullet \mathbf{Y} \leq 1, \ \forall\, i = 1, \cdots, I; & (a) \\ \mathbf{C}^T\mathbf{C} \bullet \mathbf{Y} = 0; & (b) \\ \mathcal{S}(\mathbf{Y}) \preceq \mathbf{I}_M, \ \mathcal{T}(\mathbf{Y}) \preceq \mathbf{I}_N; & (c) \\ \mathbf{Y} \in \mathbb{S}^{MN}, \ \mathbf{Y} \succeq \mathbf{0}, & (d) \end{cases} \tag{24}$$

where $\mathbb{S}^{MN}$ is the space of $MN \times MN$ symmetric matrices; $\mathbf{D}, \mathbf{B}_1, \cdots, \mathbf{B}_I$ are the $MN \times MN$ symmetric matrices corresponding to the self-adjoint linear mappings $\mathcal{D}, \mathcal{B}_1, \cdots, \mathcal{B}_I$ respectively; $\mathbf{C}$ is the $L \times MN$ matrix corresponding to the mappings $\mathcal{C}$; $\mathcal{S} : \mathbb{S}^{MN} \to \mathbb{S}^M$ is the linear mapping such that given $\mathbf{X} \in \mathbb{M}^{M\times N}$, $\mathbf{X}\mathbf{X}^T \preceq \mathbf{I}_M$ if and only if $\mathcal{S}\big((\mathrm{Vec}\,\mathbf{X})(\mathrm{Vec}\,\mathbf{X})^T\big) \preceq \mathbf{I}_M$; and $\mathcal{T} : \mathbb{S}^{MN} \to \mathbb{S}^N$ is the linear mapping such that $\mathbf{X}^T\mathbf{X} \preceq \mathbf{I}_N$ if and only if $\mathcal{T}\big((\mathrm{Vec}\,\mathbf{X})(\mathrm{Vec}\,\mathbf{X})^T\big) \preceq \mathbf{I}_N$. Refer to Section 3.1.1 of [25] for details of these notations.

By using the ellipsoid method, the solution $\widehat{\mathbf{Y}}$ to the optimization problem (24) can be obtained with an additive error $\pi > 0$ in polynomial time. That is, if θ^* is the optimal value of (24), the ellipsoid method can be used for any $\pi > 0$ to obtain a solution $\widehat{\mathbf{Y}}$ in polynomial time such that $\widehat{\mathbf{Y}}$ is feasible for (24) and satisfies $\widehat{\theta} := \mathbf{A} \bullet \widehat{\mathbf{Y}} \geq \theta^* - \pi$, where $\mathbf{A}$ is the $MN \times MN$ symmetric matrix corresponding to the self-adjoint linear mapping $\mathcal{A}$ in (23).

The solution $\widehat{\mathbf{X}} \in \mathbb{M}^{M\times N}$ to the optimization problem (23) can be achieved by using $\widehat{\mathbf{Y}}$ along with a degree of randomness. Since $\widehat{\mathbf{Y}} \succeq \mathbf{0}$, there exists a positive semidefinite matrix $\widehat{\mathbf{Y}}^{1/2} \in \mathbb{S}^{MN}$ such that $\widehat{\mathbf{Y}} = \widehat{\mathbf{Y}}^{1/2}\widehat{\mathbf{Y}}^{1/2}$. Since $\widehat{\mathbf{Y}}^{1/2}\mathbf{A}\widehat{\mathbf{Y}}^{1/2}$ is also symmetric, it has a spectral decomposition $\widehat{\mathbf{Y}}^{1/2}\mathbf{A}\widehat{\mathbf{Y}}^{1/2} = \mathbf{U}^T\Lambda\mathbf{U}$, where $\mathbf{U}$ is an $MN \times MN$ orthogonal matrix and Λ is an $MN \times MN$ diagonal matrix. Let $\boldsymbol{\xi} = (\xi_1, \xi_2, \cdots, \xi_{MN})^T$ be an MN-dimensional random vector, where ξ_n $(1 \leq n \leq MN)$ are i.i.d. with *zero* mean

and *unit* variance. The solution $\widehat{\mathbf{X}}$ is ultimately achieved via $\operatorname{Vec}\widehat{\mathbf{X}} = \widehat{\mathbf{Y}}^{1/2}\mathbf{U}^T\boldsymbol{\xi}$. Alternatively, $\widehat{\mathbf{X}}$ can be expressed as

$$\widehat{\mathbf{X}} = \sum_{i=1}^{MN} \xi_i \mathbf{Q}_i, \tag{25}$$

where $\mathbf{Q}_i \in \mathbb{R}^{M\times N}$ and $\operatorname{Vec}\mathbf{Q}_i$ is the i-th column vector of the matrix $\widehat{\mathbf{Y}}^{1/2}\mathbf{U}^T$ $(1 \le i \le MN)$. To explore the quality of solution $\widehat{\mathbf{X}}$, the following problem should be considered:

(P2) Does $\widehat{\mathbf{X}}$ act as a high-quality solution to the optimization problem (23) with a reasonable (at least larger than $1/2$) probability?

B. An Extension of Nemirovski's Conjecture

Nemirovski [24] pointed out that the aforementioned two problems P1 and P2 can be reduced to a question about the behavior of the upper bound of $\Pr\{\|\sum_k \xi_k\mathbf{A}_k\| > t\}$ and the "optimal" answer to this question can be achieved by resolving the following conjecture:

Conjecture 4.1: ([24, 25]) Let $\xi_1, \cdots, \xi_K$ be i.i.d. random variables with *zero* mean, each of which obeys either distribution supported on $[-1, 1]$ or Gaussian distribution with *unit* variance. Let $\mathbf{A}_1, \cdots, \mathbf{A}_K$ be arbitrary $M \times N$ matrices satisfying

$$\sum_{k=1}^{K} \mathbf{A}_k\mathbf{A}_k^T \preceq \mathbf{I}_M \quad \text{and} \quad \sum_{k=1}^{K} \mathbf{A}_k^T\mathbf{A}_k \preceq \mathbf{I}_N.$$

Then, whenever $t = O\big[\sqrt{\ln(M+N)}\big]$, we have

$$\mathbb{P}\left\{\left\|\sum_{k=1}^{K}\xi_k\mathbf{A}_k\right\| > t\right\} \le \theta_1 \cdot \exp(-\theta_2 \cdot t^2), \tag{26}$$

where θ_1 and θ_2 are absolute constants.

Nemirovski [24] showed that the inequality (26) is achieved when $t = O\big[(\ln(M+N))^{\frac{1}{6}}\big]$, while there is a gap between this value of t and the conjectured value $O\big[\sqrt{\ln(M+N)}\big]$. Anthony So used a non-commutative Khintchine inequality to show that when $t = O\big[\sqrt{(1+\alpha)\ln\max\{M,N\}}\big]$, for any $\alpha \ge 1/2$ (*cf.* [25]),

$$\mathbb{P}\left\{\left\|\sum_{k=1}^{K}\xi_k\mathbf{A}_k\right\| > t\right\} \le O\Big[(\max\{M,N\})^{-\alpha}\Big]. \tag{27}$$

Note that these results are built under the assumption that $\xi_1, \cdots, \xi_K$ are either Gaussian distributions or distributions supported on $[-1, 1]$. However, the assumption will not always be satisfied in practice. Therefore, we extend the content of the conjecture to the i.d. scenario, *i.e.,* whether the inequality (26) is still valid when $\xi_1, \cdots, \xi_K$ are independent i.d. random variables with *zero* mean and *unit* variance. The following theorem provides a solution to the extended version of Nemirovski's conjecture.

Theorem 4.1: Assume that $\mathbf{A}_1, \cdots, \mathbf{A}_K$ are fixed $M \times N$ matrices satisfying $\lambda_{\max}(\mathfrak{D}(\mathbf{A}_k)) \le 1$ for any $1 \le k \le K$ and denote $\rho_1 := \lambda_{\max}(\sum_k \mathfrak{D}^2(\mathbf{A}_k))$, where

$$\mathfrak{D}(\mathbf{A}) := \begin{bmatrix} \mathbf{0} & \mathbf{A} \\ \mathbf{A}^* & \mathbf{0} \end{bmatrix}.$$

Let $\xi_1, \cdots, \xi_K$ be independent i.d. random variables with the triplet (b, σ^2, ν), each of which has *zero* mean and *unit* variance. Suppose that ν has a bounded support with $R = \inf\{\alpha > 0 : \nu(\{u : |u| > \alpha\}) = 0\}$ and set $V := \int_{\mathbb{R}} |u|^2 \nu(du)$. For any $\alpha > 0$, denote

$$c_\alpha := \frac{(1+\alpha)\ln(M+N)}{\sqrt{\beta_0}} \cdot \max\{1, \sqrt{R}\} \times \max\left\{1, \sqrt{\frac{\rho_1(\sigma^2+V)}{R}}\right\}.$$

Let $\tau_\alpha := \tau(\beta_0, c_\alpha) \in (1, 2]$, where $\tau(\cdot)$ is defined in (14) and $\beta_0 = 2\ln 2 - 1$. Then, when

$$t = \left[\frac{(1+\alpha)\cdot[\rho_1(\sigma^2+V)]^{\tau_\alpha - 1}\cdot\ln(M+N)}{\beta_0 \cdot R^{\tau_\alpha - 2}}\right]^{\frac{1}{\tau_\alpha}} > \frac{\sigma^2+V}{R}, \tag{28}$$

it holds that

$$\mathbb{P}\left\{\left\|\sum_{k=1}^{K}\xi_k\mathbf{A}_k\right\| > t\right\} \le (M+N)^{-\alpha}, \quad \alpha > 0. \tag{29}$$

This theorem shows that if $\xi_1, \cdots, \xi_K$ are i.d. distributions, the probability that $\|\sum_{k=1}^K \xi_k\mathbf{A}_k\| > t$ can also be bounded by the term $(M+N)^{-\alpha}$ $(\alpha > 0)$ when $t = O\big[((1+\alpha)\ln\max\{M,N\})^{1/\tau}\big]$ $(1 < \tau \le 2)$. This solution is in accordance with So's solution (27) to the original Nemirovski conjecture up to some constant. Therefore, the discussion in Section IV-A is also valid in the setting of matrix i.d. series.

Remark 4.1: According to the tail inequality (17), when

$$t = \sqrt{\frac{(\alpha+1)\cdot[\rho_1(\sigma^2+V)]\cdot\ln(M+N)}{\beta_0}} \le \frac{\sigma^2+V}{R},$$

the result (29) still holds. However, to satisfy this condition, an assumption about the distribution of the i.d. random variable ξ_k needs to be imposed, *i.e.,* the value of R should be small enough. This will restrict the generality of the result, so we omit it here.

C. Solutions to Problems P1&P2

In this section, we will provide solutions to the aforementioned problems P1 and P2 in the i.d. scenario. By using the tail inequality (17), we first arrive at the solution to Problem P1:

Theorem 4.2: Consider the chance constrained optimization problem (19). Let $\xi_1, \cdots, \xi_K$ be independent i.d. random variables satisfying the conditions in Theorem 4.1. Denote $\rho_2 := \lambda_{\max}(\sum_k(\mathcal{A}_k'(\mathbf{x}))^2)$. For any $\epsilon \in (0, 1/2]$, let $c > 1$ satisfy that

$$2M\exp\left(-\frac{c^2\beta_0\rho_2(\sigma^2+V)}{R^2}\right) \le \epsilon. \tag{30}$$

If it holds that

$$\sum_{k=1}^{K}(\mathcal{A}_k'(\mathbf{x}))^2 \preceq \gamma\mathbf{I} \tag{31}$$

with

$$\gamma \le \gamma_2(\epsilon) := \left(\frac{\beta_0 R^{\tau_c - 2}}{\left[\rho_2(\sigma^2 + V)\right]^{\tau_c - 1} \log(\frac{2M}{\epsilon})} \right)^{\frac{1}{\tau_c}},$$

then the positive semidefinite constraint (22) is a tractable approximation of the constraint (19-b).

Note that since $\tau_c = \tau(\beta_0, c)$ takes value from the interval $(1, \frac{\log(2)}{2\log(2)-1})$ when $c > 1$, $\gamma_2(\epsilon) = O\big(\log(\frac{2M}{\epsilon})^{-1/\tau_c}\big)$ is smaller than the value $\gamma = O\big(\log(\frac{M}{\epsilon})^{-1/2}\big)$ obtained in the scenario of either the distributions with $[-1, 1]$ support or Gaussian distributions (*cf.* [25]) when the matrix size M is large.

Next, we consider the solution to Problem P2 in the matrix i.d. scenario. Consider the quadratic optimization problem (23). The following theorem proves the properties of the solution $\widehat{\mathbf{X}} = \sum_{i=1}^{MN} \xi_i \mathbf{Q}_i$ in (25).

Theorem 4.3: Following the notations in (23) and (24). Let $\xi_1, \cdots, \xi_{MN}$ be independent i.d. random variables satisfying the conditions in Theorem 4.1. Then, it holds that

i) $\mathbb{E}\big\{\widehat{\mathbf{X}} \bullet \mathcal{D}\widehat{\mathbf{X}}\big\} = \widehat{\theta}$;
ii) $\mathbb{E}\big\{\widehat{\mathbf{X}} \bullet \mathcal{B}_i\widehat{\mathbf{X}}\big\} \le 1,\ \forall i = 1, \cdots, I$;
iii) $\mathcal{C}\widehat{\mathbf{X}} = 0$;
iv) $\mathbb{E}\big\{\widehat{\mathbf{X}}\widehat{\mathbf{X}}^T\big\} = \mathbf{I}_M$ and $\mathbb{E}\big\{\widehat{\mathbf{X}}^T\widehat{\mathbf{X}}\big\} = \mathbf{I}_N$.

Its proof is similar to the proof of Proposition 1 in [25], so we omit it here.

This theorem shows that the matrix i.d. series $\widehat{\mathbf{X}} = \sum_{j=1}^{MN} \xi_j \mathbf{Q}_j$ satisfies the constraints of the original optimization problem (23) when taking expectation. It remains to justify whether $\widehat{\mathbf{X}}$ can also satisfy the constraints (23-a) and (23-c) with reasonable probability (at least larger than $1/2$).

Theorem 4.4: Assume that $\xi_1, \cdots, \xi_{MN}$ are independent i.d. random variables satisfying the conditions in Theorem 4.1. Let $\mathbf{B}_i' = \mathbf{U}\widehat{\mathbf{Y}}^{1/2}\mathbf{B}_i\widehat{\mathbf{Y}}^{1/2}\mathbf{U}^T$ $(i = 1, \cdots, I)$ and denote by $\mathrm{col}_j[(\mathbf{B}_i')^{1/2}]$ the matrix whose j-th column is the j-th column of the matrix $(\mathbf{B}_i')^{1/2}$ and the other entries are all *zero* $(j = 1, \cdots, MN)$. Denote $\rho_3 := \lambda_{\max}(\sum_{j=1}^{MN} \mathbf{Q}_j^2)$ and $\rho_4^{(i)} := \lambda_{\max}\big(\sum_{j=1}^{MN} (\mathrm{col}_j[(\mathbf{B}_i')^{1/2}])^2\big)$. Then, with probability at least $1/2$, it holds that

$$\|\widehat{\mathbf{X}}\| \le \left[\frac{3[\rho_3(\sigma^2 + V)]^{\tau_2 - 1} \cdot \ln(M + N)}{\beta_0 \cdot R^{\tau_2 - 2}} \right]^{\frac{1}{\tau_2}}, \qquad (32)$$

and for any $1 \le i \le I$

$$\widehat{\mathbf{X}} \bullet \mathcal{B}_i\widehat{\mathbf{X}} \le \left[\frac{3[\rho_4^{(i)}(\sigma^2 + V)]^{\tau_2 - 1} \cdot \ln(M + N)}{\beta_0 \cdot R^{\tau_2 - 2}} \right]^{\frac{2}{\tau_2}}. \qquad (33)$$

This theorem implies that

$$\overline{\mathbf{X}} := \widehat{\mathbf{X}} \cdot \left[\frac{3[\rho_*(\sigma^2 + V)]^{\tau_2 - 1} \cdot \ln(M + N)}{\beta_0 \cdot R^{\tau_2 - 2}} \right]^{\frac{-1}{\tau_2}}$$

is feasible to the quadratic optimization problem (23) with a probability larger than $1/2$, where $\rho_* = \max\{\rho_3, \rho_4^{(1)}, \rho_4^{(2)}, \cdots, \rho_4^{(I)}\}$. It thus also provides a solution to Problem P2.

V. Conclusion

The class of i.d. distributions is large and includes important probability distributions, such as Gaussian and Poisson distributions, that are widely used in several fields. To the best of our knowledge, however, little work has been done on random matrix theory with respect to i.d. distributions. In this paper, we are mainly concerned with the tail inequalities of the largest eigenvalue of a matrix i.d. series, and our results encompass Tropp's work [1] on matrix Gaussian series as a special case. Our proof strategy is as follows. We first relax the Bennett-type result (6) into a Bernstein-type result (10) by replacing $Q(s)$ with $B(s)$ or $T(s)$ (8). Subsequently, we present an upper bound on the expectation $\mathbb{E}\big\|\sum_k \xi_k \mathbf{A}_k\big\|$, which is looser than the bound for the Gaussian case (*cf.* Inequality (4.9) of [1]) because of the existence of compound Poisson components in the i.d. distribution (*cf.* the Lévy-Itô decomposition).

Since the function $B(s)$ does not bound $Q(s)$ from below sufficiently tightly (*cf.* Fig. 4), we develop a new lower-bound function $H_P(s)$ to bound $Q(s)$ from below on a bounded domain $s \in (0, c]$, where the partition $P = \{S_0, S_1, \cdots, S_N\}$ is an ordered sequence such that $1 = S_0 < S_1 < \cdots < S_N = c$ for any given $c \in (1, +\infty)$. Although $H_P(s)$ is a piecewise function, its computational cost is low because all sub-functions of $H_P(s)$ are uniformly expressed in the form $\beta_0 \cdot s_n^{\tau}$, where $\beta_0 = 2\log 2 - 1$ and $\tau_n = \tau(\beta_0, S_n)$ $(n = 1, 2, \cdots, N)$. Based on $H_P(s)$, we obtain another tail inequality for matrix i.d. series that is tighter than the Bernstein-type result given in (10) when $\frac{Rt}{\rho(\sigma^2+V)} > 0.8831$ and provides a tighter upper bound on $\lambda_{\max}\big(\sum_k \xi_k \mathbf{A}_k\big)$ when the matrix dimension d is high. Our results concerning the functions $Q(s)$ and $H_P(s)$ are also applicable for any Bennett-type concentration inequality that involves the function $Q(s)$.

In addition, we study the application of random i.d. series in several optimization problems including 1) the safe tractable approximation of chance constrained linear matrix inequalities, and 2) the quality of the semidefinite relaxation of a general non-convex quadratic optimization problem with orthogonality constraints, which covers two well-studied optimization problems as special cases: the Procrustes problem and the quadratic assignment problem. These two problems have been extensively studied in [24, 25] under the assumption that $\{\xi_k\}$ are sub-Gaussian, whereas in reality this assumption will not always be satisfied. We are able to extend the feasibility of the findings in [24, 25] to the case in which $\{\xi_k\}$ are i.d. distributions.

Since the tail inequalities considered in this paper depend on the matrix dimension, they will become loose in the high-dimensional case [12]. Similar to the results obtained in existing works, these inequalities can be improved by introducing the concept of effective dimension [15] or intrinsic dimension [14]. In our future work, we will also consider the extension of these results to the infinite-dimensional case.

Appendix A
Lévy Measure

Before introducing the Lévy measure, we first present a discussion of Lévy processes. For further details, the reader is

referred to [27].

Definition A.1 (Lévy Process): A process $\mathcal{X} = \{X_t : t \geq 0\}$, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, is said to be a Lévy process if it has the following properties:

1) The paths of $\mathcal{X}$ are $\mathbb{P}$-almost surely right continuous with left limits.
2) $\mathbb{P}(X_0 = 0) = 1$.
3) For $0 \leq s \leq t$, $X_t - X_s$ is equal in distribution to X_{t-s}.
4) For $0 \leq s \leq t$, $X_t - X_s$ is independent of $\{X_u : u \leq s\}$.

Given a Lévy process $\{X_t : t \geq 0\}$, consider the jump process $\Delta\mathcal{X} := \{\Delta X_t\}_{0 \leq t \leq T}$, that is,

$$\Delta X_t = X_t - X_{t^-}, \quad \forall\, 0 \leq t \leq T,$$

where $X_{t^-} := \lim_{s \to t^-} X_s$. It follows Definition A.1 that for any fixed $t > 0$, $\Delta X_t = 0$ almost surely.

Moreover, given a set $A \in \mathcal{B}(\mathbb{R}/\{0\})$ such that $0 \notin \overline{A}$, let the random measure of the jumps be defined as

$$\begin{aligned}\mu(\boldsymbol{\omega}; t, A) :=& \#\{0 \leq s \leq t; \Delta X_s(\omega_s) \in A\} \\ =& \sum_{s \leq t} 1_A(\Delta X_s(\omega_s)), \qquad 0 \leq t \leq T,\end{aligned}$$

where $\boldsymbol{\omega}$ denotes joint probability events in the time interval $[0, t]$ and ω_s denotes events related to the s-time distribution of the Lévy process $\{X_t : t \geq 0\}$. As defined above, the measure $\mu(\boldsymbol{\omega}; t, A)$ counts the number of jumps of a size included in A up to time t in the process $\{X_t : t \geq 0\}$.

The Lévy measure is finally defined as

$$\begin{aligned}\nu(A) :=& \mathbb{E}\big\{\mu(\boldsymbol{\omega}; 1, A)\big\} = \mathbb{E}\left\{\sum_{s \leq 1} 1_A(\Delta X_s(\omega_s))\right\} \\ =& \sum_{s \leq 1} \mathbb{E}\big\{1_A(\Delta X_s(\omega_s))\big\} \quad \text{(jumps are independent)} \\ =& \sum_{s \leq 1} \mathbb{E}_s\big\{1_A(\Delta X_s(\omega_s))\big\},\end{aligned}$$

where the expectations $\mathbb{E}$ and $\mathbb{E}_s$ are taken w.r.t. $\boldsymbol{\omega}$ and ω_s, respectively. The Lévy measure describes the expected number of jumps of a certain height (belonging to A) in a time interval of *unit* length.

APPENDIX B
PROOFS OF THE MAIN RESULTS

Here, we prove Lemma 2.1, Theorem 3.1, Corollary 3.1, Theorem 3.2 and Theorem 4.1.

A. Proof of Lemma 2.1

Let $\psi(\theta) : \mathbb{R} \to \mathbb{C}$ denote the characteristic function of the i.d. random variable $\xi \in \mathbb{R}$ with the triplet (b, σ^2, ν). Let $(\xi_0, \xi_0'), (\xi_1, \xi_1') \in \mathbb{R} \times \mathbb{R}$ be i.d. vectors with the characteristic functions $\psi_0(\theta, \theta') = \psi(\theta) \cdot \psi(\theta')$ and $\psi_1(\theta, \theta') = \psi(\theta + \theta')$ $(\theta, \theta' \in \mathbb{R})$ respectively. For any $0 \leq r \leq 1$, let (ξ_r, ξ_r') be a random vector with the characteristic function[4]

$$\begin{aligned}\psi_r(\theta, \theta') :=& \big[\psi_0(\theta, \theta')\big]^{1-r} \cdot \big[\psi_1(\theta, \theta')\big]^r \\ =& \big[\psi(\theta) \cdot \psi(\theta')\big]^{1-r} \big[\psi(\theta + \theta')\big]^r. \qquad (34)\end{aligned}$$

Remark B.1: Here, we justify why $\psi_r(\theta, \theta')$ is a characteristic function. Recalling Definition 2.1, we have $\phi(\theta) = \log \psi(\theta)$, thus $r\phi(\theta) = \log\big[\psi(\theta)\big]^r$ for any $0 \leq r \leq 1$. It follows from Theorem 2.1 that $\big[\psi(\theta)\big]^r$ is the characteristic function of the i.d. random variable with the triplet $(rb, r\sigma^2, r\nu)$. Since the product of a finite number of characteristic functions is also a characteristic function, the term $\psi_r(\theta, \theta')$ is a characteristic function.

To prove Lemma 2.1, we first need the following lemma, which is the one-dimensional case of Proposition 2 of [35].

Lemma B.1: Let ξ be an i.d. random variable with the triplet (b, σ^2, ν). If $f, g : \mathbb{R} \to \mathbb{R}$ are differentiable functions such that $\mathbb{E}|f(\xi)|, \mathbb{E}|g(\xi)|, \mathbb{E}|f(\xi)g(\xi)| < \infty$, then

$$\begin{aligned}\mathbb{E}f(\xi)g(\xi) - \mathbb{E}f(\xi)\mathbb{E}g(\xi) = \int_0^1 \mathbb{E}_r \Big\{&\sigma^2 \nabla f(\xi_r) \cdot \nabla g(\xi_r) \\ &+ \int_{\mathbb{R}} \big(f(\xi_r + u) - f(\xi_r)\big)\big(g(\xi_r' + u) - g(\xi_r')\big)\nu(du)\Big\} dr,\end{aligned}$$

where the expectation $\mathbb{E}_r$ is taken on the joint distribution of (ξ_r, ξ_r') and ∇ is the derivative notation.

The expectation $\mathbb{E}_r$ has the following properties:

Lemma B.2: If $\psi_r(-i\theta, -i\theta')$ exists for any $r \in [0, 1]$, it holds that

$$\mathbb{E}_r\big\{\mathrm{e}^{\theta\xi_r'}\big\} = \mathbb{E}\big\{\mathrm{e}^{\theta\xi}\big\}. \qquad (35)$$

Proof of Lemma B.2. According to (34), for any $r \in [0, 1]$, we arrive at

$$\begin{aligned}\mathbb{E}_r\big\{\mathrm{e}^{\theta\xi_r'}\big\} =& \psi_r(-i0, -i\theta) \\ =& \big[\psi(-i0) \cdot \psi(-i\theta)\big]^{1-r}\big[\psi(-i0 - i\theta)\big]^r \\ =& \big[\psi(-i\theta)\big]^{1-r}\big[\psi(-i\theta)\big]^r \\ =& \psi(-i\theta) = \mathbb{E}\big\{\mathrm{e}^{\theta\xi}\big\}. \qquad (36)\end{aligned}$$

This completes the proof. ■

Lemma 2.1 can be proven by using the techniques presented in Houdré's work [31].

Proof of Lemma 2.1. As stated in Theorem 25.3 of [36], since the function e^y $(y \in \mathbb{R})$ is submultiplicative, it holds that

$$\begin{aligned}\Omega :=& \Big\{s \geq 0 : \mathbb{E}\mathrm{e}^{s|\xi|} < +\infty\Big\} \\ =& \left\{s \in \mathbb{R} : \int_{|u|>1} \mathrm{e}^{s|u|}\nu(du) < +\infty\right\}.\end{aligned}$$

[4]Recalling Definition 2.1, we have $\phi(\theta) = \log \psi(\theta)$, thus $r\phi(\theta) = \log\big[\psi(\theta)\big]^r$ for any $0 \leq r \leq 1$. It follows from Theorem 2.1 that $\big[\psi(\theta)\big]^r$ is the characteristic function of an i.d. random variable with the triplet $(rb, r\sigma^2, r\nu)$. Since the product of a finite number of characteristic functions is also a characteristic function, the function $\psi_r(\theta, \theta')$ is ultimately proven to be a characteristic function.

Since $0 < \mathrm{e}^{s|u|} - s|u| - 1 < \mathrm{e}^{s|u|}$, it follows the definition of the Lévy measure ν (*cf.* Definition 2.2) that

$$\Omega = \left\{ s \geq 0 : \int_{|u|>1} \left(\mathrm{e}^{s|u|} - s|u| - 1 \right) \nu(du) < +\infty \right\}.$$

Based on the convexity of the exponential function, the set Ω is an interval of $\mathbb{R}$ and contains *zero*, but it cannot degenerate to $\{0\}$. We adopt the notation $\Omega = [0, M]$ with

$$M = \sup \left\{ s \geq 0 : \int_{|u|>1} \mathrm{e}^{s|u|} \nu(du) < +\infty \right\}.$$

Thus, the following discussion is valid.

By Lemma B.1, we have

$$\begin{aligned}
&\mathbb{E}\{\xi \cdot \mathrm{e}^{s\xi}\} - \mathbb{E}\xi \cdot \mathbb{E}\mathrm{e}^{s\xi} \\
=&\int_0^1 \mathbb{E}_r \left\{ \sigma^2 \cdot \frac{\mathrm{d}\mathrm{e}^{s\xi_r'}}{\mathrm{d}\xi_r'} \right. \\
&\left. + \int_{\mathbb{R}} (\xi_r + u - \xi_r)\big(\mathrm{e}^{s(\xi_r'+u)} - \mathrm{e}^{s\xi_r'}\big)\nu(du) \right\} dr \\
=&\int_0^1 \mathbb{E}_r \left\{ s\mathrm{e}^{s\xi_r'}\sigma^2 + \mathrm{e}^{s\xi_r'} \int_{\mathbb{R}} u\big(\mathrm{e}^{su} - 1\big)\nu(du) \right\} dr \\
\leq&\left(\sigma^2 s + \int_{\mathbb{R}} |u|\big(\mathrm{e}^{s|u|} - 1\big)\nu(du) \right) \cdot \int_0^1 \mathbb{E}_r\{\mathrm{e}^{s\xi_r'}\} dr \\
=&\left(\sigma^2 s + \int_{\mathbb{R}} |u|\big(\mathrm{e}^{s|u|} - 1\big)\nu(du) \right) \cdot \mathbb{E}\{\mathrm{e}^{s\xi}\}. \qquad (37)
\end{aligned}$$

The last equality is derived from Lemma B.2.

Let $L(s) := \mathbb{E}\mathrm{e}^{s\xi'}$. It follows from $\mathbb{E}\xi = 0$ that

$$\frac{\mathrm{d}L(s)}{\mathrm{d}s}\frac{1}{L(s)} = \frac{\mathbb{E}\xi\mathrm{e}^{s\xi}}{\mathbb{E}\mathrm{e}^{s\xi}} \leq \sigma^2 s + \int_{\mathbb{R}} |u| \left(\mathrm{e}^{s|u|} - 1 \right) \nu(du).$$

Therefore, we have

$$\begin{aligned}
&\int_0^\theta \frac{\mathrm{d}L(s)}{\mathrm{d}s}\frac{1}{L(s)} ds \\
\leq&\int_0^\theta \left(\sigma^2 s + \int_{\mathbb{R}} |u| \left(\mathrm{e}^{s|u|} - 1 \right) \nu(du) \right) ds,
\end{aligned}$$

thus

$$\log \mathbb{E}\mathrm{e}^{s\xi}\Big|_0^\theta \leq \frac{\sigma^2\theta^2}{2} + \int_{\mathbb{R}} \left(\mathrm{e}^{\theta|u|} - \theta|u| - 1 \right) \nu(du). \qquad (38)$$

From the proof of Lemma 6.7 in [1], we obtain the following inequality:

$$\frac{\mathrm{e}^{\lambda\theta|u|} - \lambda\theta|u| - 1}{\lambda^2} \leq \mathrm{e}^{\theta|u|} - \theta|u| - 1, \ \ \forall\, \lambda \leq 1. \qquad (39)$$

By combining (38) and (39), we have for any $\lambda \leq 1$,

$$\mathbb{E}\mathrm{e}^{\lambda\theta\xi} \leq \exp\left(\frac{\sigma^2\theta^2\lambda^2}{2} + \lambda^2 \int_{\mathbb{R}} \left(\mathrm{e}^{\theta|u|} - \theta|u| - 1 \right) \nu(du) \right).$$

Given a self-adjoint matrix $\mathbf{A}$ with $\lambda_{\max}(\mathbf{A}) \leq 1$, it follows the transfer rule that

$$\mathbb{E}\mathrm{e}^{\xi\theta\mathbf{A}} \preceq \mathrm{e}^{\Phi(\theta)\cdot\mathbf{A}^2}, \qquad (40)$$

where for any $0 < \theta < M$,

$$\Phi(\theta) := \frac{\sigma^2\theta^2}{2} + \int_{\mathbb{R}} \left(\mathrm{e}^{\theta|u|} - \theta|u| - 1 \right) \nu(du). \qquad (41)$$

This completes the proof. ■

B. Proof of Theorem 3.1

Proof of Theorem 3.1: Let $\rho := \lambda_{\max}\big(\sum_k \mathbf{A}_k^2\big)$. It follows from Lemma 2.1 that, for any $t > 0$,

$$\begin{aligned}
&\mathbb{P}\left\{ \lambda_{\max}\big(\sum_k \xi_k\mathbf{A}_k\big) > t \right\} \\
\leq&\mathrm{e}^{-\theta t} \cdot \mathrm{tr}\exp\left(\sum_k \log \mathbb{E}\mathrm{e}^{\theta t_k \mathbf{A}_k} \right) \\
\leq&\mathrm{e}^{-\theta t} \cdot \mathrm{tr}\exp\left(\Phi(\theta) \cdot \sum_k \mathbf{A}_k^2 \right) \\
\leq&\mathrm{e}^{-\theta t} \cdot d \cdot \lambda_{\max}\left(\exp\left(\Phi(\theta) \cdot \sum_k \mathbf{A}_k^2 \right) \right) \\
=&d \cdot \exp\left(-\theta t + \Phi(\theta) \cdot \lambda_{\max}\left(\sum_k \mathbf{A}_k^2 \right) \right) \\
=&d \cdot \exp\left(-\theta t + \Phi(\theta) \cdot \rho \right), \qquad (42)
\end{aligned}$$

where the first inequality follows from Theorem 3.6 of [1].

By (4), since $\mathbb{E}\mathrm{e}^{\theta\xi} < +\infty$ for all $0 < \theta < M$, $\Phi(\theta)$ is infinitely differentiable on $(0, M)$, with

$$\Phi'(\theta) := \alpha(\theta) = \sigma^2\theta + \int_{\mathbb{R}} |u|\big(\mathrm{e}^{\theta|u|} - 1\big)\nu(du) > 0, \qquad (43)$$

and

$$\Phi''(\theta) = \sigma^2 + \int_{\mathbb{R}} |u|^2 \mathrm{e}^{\theta|u|}\nu(du) > 0. \qquad (44)$$

Then, we minimize the right-hand side of (42) w.r.t. θ. According to (43) and (44), for any $0 < t < \frac{\alpha(M^-)}{\rho}$, $\min_{0<\theta<M}\{\rho\cdot\Phi(\theta) - \theta\cdot t\}$ is achieved when $\theta = \alpha^{-1}(t/\rho)$. Since $\Phi(0) = \alpha(0) = \alpha^{-1}(0) = 0$, we have

$$\begin{aligned}
\Phi\left(\alpha^{-1}(t/\rho)\right) =& \int_0^{\alpha^{-1}(t/\rho)} \alpha(s)\, ds \\
=& \int_0^{t/\rho} s\, d\alpha^{-1}(s) \\
=&(t/\rho)\cdot\alpha^{-1}(t/\rho) - \int_0^{t/\rho} \alpha^{-1}(s)\, ds. \qquad (45)
\end{aligned}$$

Thus, for any $0 < t < \frac{\alpha(M^-)}{\rho}$,

$$\begin{aligned}
\min_{0<\theta<M}\{\rho\cdot\Phi(\theta) - \theta\cdot t\} =& \rho\cdot\Phi\left(\alpha^{-1}(t/\rho)\right) - t\cdot\alpha^{-1}(t/\rho) \\
=& -\rho\cdot\int_0^{t/\rho} \alpha^{-1}(s) ds.
\end{aligned}$$

This completes the proof. ■

C. Proof of Corollary 3.1

Proof of Corollary 3.1: Since the support is $\mathrm{supp}(\nu) \subseteq [-R, R]$, it holds that $\mathbb{E}\mathrm{e}^{\theta|\xi|} < +\infty$ for any $\theta > 0$. Thus, we

have

$$\begin{aligned}\alpha(\theta) =&\sigma^2\theta+\int_{\mathbb{R}}|u|\big(\mathrm{e}^{\theta|u|}-1\big)\nu(du)\\ =&\sigma^2\theta+\int_{|u|\le R}|u|^2\left(\sum_{k=1}^{\infty}\frac{\theta^k|u|^{k-1}}{k!}\right)\nu(du)\\ \le&\sigma^2\theta+\int_{|u|\le R}|u|^2\left(\sum_{k=1}^{\infty}\frac{\theta^kR^{k-1}}{k!}\right)\nu(du)\\ =&\sigma^2\theta+V\left(\frac{\mathrm{e}^{\theta R}-1}{R}\right)\le(\sigma^2+V)\left(\frac{\mathrm{e}^{\theta R}-1}{R}\right).\end{aligned}\tag{46}$$

Note that if the strictly increasing functions $\alpha,\beta:\mathbb{R}^+\to\mathbb{R}^+$ satisfy $\alpha(s)\le\beta(s)$ for all $s\ge0$, then their inverse functions satisfy $\beta^{-1}(s)\le\alpha^{-1}(s)$ for all $s\ge0$. As shown in (43) and (44), $\alpha(s)$ is an increasing function, thus $\alpha^{-1}(s)$ is also an increasing function. By combining (5) and (46), we obtain, for any $t>0$,

$$\begin{aligned}&\mathbb{P}\left\{\lambda_{\max}\left(\sum_k\xi_k\mathbf{A}_k\right)>t\right\}\\ \le&d\cdot\exp\left(-\rho\cdot\int_0^{t/\rho}\alpha^{-1}(s)ds\right)\\ \le&d\cdot\exp\left(-\rho\cdot\int_0^{t/\rho}\frac{1}{R}\cdot\log\left(1+\frac{Rs}{\sigma^2+V}\right)ds\right)\\ =&d\cdot\exp\left(-\frac{\rho(\sigma^2+V)}{R^2}\cdot Q\left(\frac{Rt}{\rho(\sigma^2+V)}\right)\right),\end{aligned}$$

where $Q(s):=(1+s)\cdot\log(1+s)-s$. This completes the proof. ■

D. Proof of Theorem 3.2

Proof of Theorem 3.2: Based on the special partition $S_0=1$ and $S_1=+\infty$, we arrive at the following tail inequality for a matrix i.d. series: for any $t\in(0,+\infty)/\{1\}$,

$$\begin{aligned}&\mathbb{P}\left\{\Big\|\sum_k\xi_k\mathbf{A}_k\Big\|>t\right\}\\ \le&\begin{cases}2d\cdot\exp\left(-\frac{\beta_0t}{R}\right), & \text{if } t>\frac{\rho(\sigma^2+V)}{R};\\ 2d\cdot\exp\left(-\frac{\beta_0t^2}{\rho(\sigma^2+V)}\right), & \text{if } 0<t\le\frac{\rho(\sigma^2+V)}{R},\end{cases}\end{aligned}\tag{47}$$

where $\beta_0=2\log2-1$. Since $x+\mathrm{e}^{-x}\le1+x^2/2$ $(x>0)$, we have

$$\begin{aligned}\mathbb{E}\Big\|\sum_k\xi_k\mathbf{A}_k\Big\|=&\int_0^{+\infty}\mathbb{P}\left\{\Big\|\sum_k\xi_k\mathbf{A}_k\Big\|>t\right\}dt\\ \le&\beta_0^{-1}\cdot\log\left(2d\cdot\mathrm{e}^{\frac{\rho(\sigma^2+V)}{R}}\right)\\ &+2d\cdot\int_{\beta_0^{-1}\log\left(2d\cdot\mathrm{e}^{\frac{\rho(\sigma^2+V)}{R}}\right)}^{+\infty}\mathrm{e}^{-\beta_0t}d\xi\\ =&\beta_0^{-1}\cdot\log\left(2d\cdot\mathrm{e}^{\frac{\rho(\sigma^2+V)}{R}}\right)+\beta_0^{-1}\cdot\mathrm{e}^{-\frac{\rho(\sigma^2+V)}{R}}\\ =&\beta_0^{-1}\cdot\log\left(2d\cdot\mathrm{e}^{\frac{\rho(\sigma^2+V)}{R}+\mathrm{e}^{-\frac{\rho(\sigma^2+V)}{R}}}\right)\\ \le&\beta_0^{-1}\cdot\log\left(2d\cdot\mathrm{e}^{1+\frac{\rho^2(\sigma^2+V)^2}{2R^2}}\right).\end{aligned}$$

This completes the proof. ■

E. Proof of Theorem 4.1

Proof of Theorem 4.1: First, if t satisfies the condition (28), we have

$$\begin{aligned}t=&\left[(\alpha+1)\cdot\ln(M+N)\cdot\frac{[\rho_1(\sigma^2+V)]^{\tau_\alpha-1}}{\beta_0\cdot R^{\tau_\alpha-2}}\right]^{\frac{1}{\tau_\alpha}}\\ =&\left[\frac{(\alpha+1)\cdot R\cdot\ln(M+N)}{\beta_0}\right]^{\frac{1}{\tau_\alpha}}\cdot\frac{[\rho_1(\sigma^2+V)]^{1-\frac{1}{\tau_\alpha}}}{R^{1-\frac{1}{\tau_\alpha}}}.\end{aligned}$$

Since it follows from (14) and (15) that $1<\tau_\alpha\le2$ for any $\alpha>0$, we arrive at

$$t<\frac{(\alpha+1)\ln(M+N)}{\sqrt{\beta_0}}\cdot\max\{1,\sqrt{R}\}\cdot\max\left\{1,\sqrt{\frac{\sigma^2+V}{R}}\right\},$$

which suggests that $t<c_\alpha$ $(\forall\,\alpha>0)$. By using the dilation method (*cf.* Section 2.6 of [1]), we then have

$$\left\|\sum_{k=1}^K\xi_k\mathbf{A}_k\right\|=\lambda_{\max}\left(\sum_{k=1}^K\xi_k\mathfrak{D}(\mathbf{A}_k)\right),\tag{48}$$

where

$$\mathfrak{D}(\mathbf{A}):=\begin{bmatrix}\mathbf{0}&\mathbf{A}\\ \mathbf{A}^*&\mathbf{0}\end{bmatrix}.$$

Note that $s_{\max}(\mathbf{A}_k)=\lambda_{\max}(\mathfrak{D}(\mathbf{A}_k))\le1$ for all $k=1,2,\cdots,K$. According to (17), we then arrive at

$$\begin{aligned}&\mathbb{P}\left\{\left\|\sum_k\xi_k\mathbf{A}_k\right\|>t\right\}\\ =&\mathbb{P}\left\{\lambda_{\max}\left(\sum_k\xi_k\mathfrak{D}(\mathbf{A}_k)\right)>t\right\}\\ \le&\begin{cases}(M+N)\cdot\exp\left(-\frac{\beta_0}{\rho_1(\sigma^2+V)}\cdot t^2\right),\\ \qquad\qquad\text{if } 0<\frac{Rt}{\rho_1(\sigma^2+V)}\le1;\\ (M+N)\cdot\exp\left(-\frac{\beta_0\cdot R^{\tau_{c_\alpha}-2}}{[\rho_1(\sigma^2+V)]^{\tau_{c_\alpha}-1}}\cdot t^{\tau_{c_\alpha}}\right),\\ \qquad\qquad\text{if } 1<\frac{Rt}{\rho_1(\sigma^2+V)}\le c_\alpha,\end{cases}\end{aligned}\tag{49}$$

Substituting (28) into the last inequality of (49) leads to the result (29). This completes the proof. ■

F. Proof of Theorem 4.2

Proof of Theorem 4.2: According to (31), it holds that $\lambda_{\max}\big(\mathcal{A}_k'(\mathbf{x})/\gamma\big)\le1$. We will consider two cases respectively: 1) $\gamma\ge\frac{R}{\rho_2(\sigma^2+V)}$; and 2) $\frac{R}{c\rho_2(\sigma^2+V)}\le\gamma<\frac{R}{\rho_2(\sigma^2+V)}$ for an arbitrary $c>1$.

When $\gamma\ge\frac{R}{\rho(\sigma^2+V)}$, it follows from (6) that

$$\begin{aligned}&\mathbb{P}\left\{\left\|\sum_k\xi_k\Big(\frac{1}{\gamma}\mathcal{A}_k'(\mathbf{x})\Big)\right\|>\frac{1}{\gamma}\right\}\\ \le&2M\exp\left\{-\frac{\beta_0}{\rho_2(\sigma^2+V)\gamma^2}\right\}.\end{aligned}\tag{50}$$

Given an $\epsilon\in(0,1/2)$, if it satisfies that $2M\exp\big\{-\frac{\beta_0}{\rho(\sigma^2+V)\gamma^2}\big\}\le\epsilon$, then the choice of γ should satisfy that

$$\gamma\le\gamma_1(\epsilon):=\sqrt{\frac{\beta_0}{\rho_2(\sigma^2+V)\log\big(\frac{2M}{\epsilon}\big)}},\tag{51}$$

and meanwhile guarantee that $\frac{R}{\rho_2(\sigma^2+V)} \leq \gamma_1(\epsilon)$, which means that

$$\epsilon > 2M \cdot \exp\left(-\frac{\beta_0 \rho_2(\sigma^2+V)}{R^2}\right).$$

This relation is only valid when R is sufficiently large, so the case of $\gamma \geq \frac{R}{\rho_2(\sigma^2+V)}$ is not friendly enough to facilitate the optimization problem. We will omit this case

When $\frac{R}{c\rho_2(\sigma^2+V)} \leq \gamma < \frac{R}{\rho_2(\sigma^2+V)}$ for an arbitrary $c > 1$, it also follows from (17) that

$$\begin{aligned}&\mathbb{P}\left\{\left\|\sum_k \xi_k\left(\frac{1}{\gamma}\mathcal{A}_k'(\mathbf{x})\right)\right\| > \frac{1}{\gamma}\right\}\\ \leq& 2M\exp\left\{-\frac{\beta_0 R^{\tau_c-2}}{\left[\rho_2(\sigma^2+V)\right]^{\tau_c-1}\gamma^{\tau_c}}\right\}. \end{aligned} \tag{52}$$

For any $\epsilon \in (0, 1/2)$, if the right-hand side of (52) can be bounded by the constant ϵ, the choice of γ should satisfy the following condition:

$$\gamma \leq \gamma_2(\epsilon) := \left(\frac{\beta_0 R^{\tau_c-2}}{\left[\rho_2(\sigma^2+V)\right]^{\tau_c-1}\log(\frac{2M}{\epsilon})}\right)^{\frac{1}{\tau_c}}.$$

It is clear that when

$$\begin{aligned}&2M\exp\left(-\frac{c^2\beta_0\rho_2(\sigma^2+V)}{R^2}\right)\\ &\leq \epsilon \leq 2M\exp\left(-\frac{\beta_0\rho_2(\sigma^2+V)}{R^2}\right),\end{aligned} \tag{53}$$

it holds that $\frac{R}{c\rho_2(\sigma^2+V)} \leq \gamma \leq \gamma_2(\epsilon) < \frac{R}{\rho_2(\sigma^2+V)}$. The first inequality of (53) holds by setting appropriate $c > 1$ and the second inequality holds when ϵ is small enough. Therefore, the validity of the inequality (53) is guaranteed. We then arrive at

$$\begin{aligned}&\mathbb{P}\left\{\sum_k \xi_k \mathcal{A}_k'(\mathbf{x}) \preceq \mathbf{I}\right\} = \mathbb{P}\left\{\left\|\sum_k \xi_k \mathcal{A}_k'(\mathbf{x})\right\| \leq 1\right\}\\ =&\mathbb{P}\left\{\left\|\sum_k \xi_k\left(\frac{1}{\gamma}\mathcal{A}_k'(\mathbf{x})\right)\right\| \leq \frac{1}{\gamma}\right\} > 1-\epsilon.\end{aligned} \tag{54}$$

This completes the proof. ■

G. Proof of Theorem 4.4

Proof of Theorem 4.4: By setting $\alpha = 2$, it follows from Theorem 4.1 that with probability at least $1/4$,

$$\|\widehat{\mathbf{X}}\| \leq \left[\frac{3[\rho_3(\sigma^2+V)]^{\tau_2-1}\cdot\ln(M+N)}{\beta_0\cdot R^{\tau_2-2}}\right]^{\frac{1}{\tau_\alpha}}. \tag{55}$$

For any $1 \leq i \leq I$, we have

$$\widehat{\mathbf{X}} \bullet \mathcal{B}_i\widehat{\mathbf{X}} = \mathbf{B}_i \bullet \widehat{\mathbf{Y}}^{1/2}\mathbf{U}^T\boldsymbol{\xi}\boldsymbol{\xi}^T\mathbf{U}\widehat{\mathbf{Y}}^{1/2} = \boldsymbol{\xi}^T\mathbf{B}_i'\boldsymbol{\xi},$$

where $\mathbf{B}_i' = \mathbf{U}\widehat{\mathbf{Y}}^{1/2}\mathbf{B}_i\widehat{\mathbf{Y}}^{1/2}\mathbf{U}^T \succeq \mathbf{0}$ because $\mathbf{B}_i \succeq \mathbf{0}$. Then, we can equivalently rewrite

$$\widehat{\mathbf{X}} \bullet \mathcal{B}_i\widehat{\mathbf{X}} = \|(\mathbf{B}_i')^{1/2}\boldsymbol{\xi}\|^2 = \left\|\sum_{j=1}^{MN}\xi_j \mathrm{col}_j[(\mathbf{B}_i')^{1/2}]\right\|^2.$$

According to Theorem 4.1, for any $i = 1, 2, \cdots, I$, we have with probability at least $1/4$,

$$\widehat{\mathbf{X}} \bullet \mathcal{B}_i\widehat{\mathbf{X}} \leq \left[\frac{3[\rho_4^{(i)}(\sigma^2+V)]^{\tau_2-1}\cdot\ln(M+N)}{\beta_0\cdot R^{\tau_2-2}}\right]^{\frac{2}{\tau_2}}. \tag{56}$$

Therefore, both of the inequalities (55) and (56) are valid with probability at least $1-(1/4+1/4) = 1/2$. This completes the proof. ■

REFERENCES

[1] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Foundations of Computational Mathematics*, vol. 12, no. 4, pp. 389–434, 2012.

[2] J. Andersson and J. O. Stromberg, "On the theorem of uniform recovery of random sampling matrices," *IEEE Transactions on Information Theory*, vol. 60, no. 3, pp. 1700–1710, 2014.

[3] S. Dirksen, G. Lecue, and H. Rauhut, "On the gap between restricted isometry properties and sparse recovery conditions," *IEEE Transactions on Information Theory*, vol. 64, no. 8, pp. 5478–5487, 2016.

[4] M. Vehkaperä, Y. Kabashima, and S. Chatterjee, "Analysis of regularized ls reconstruction and random matrix ensembles in compressed sensing," *IEEE Transactions on Information Theory*, vol. 62, no. 4, pp. 2100–2124, 2016.

[5] L. Wei, R. A. Pitaval, J. Corander, and O. Tirkkonen, "From random matrix theory to coding theory: Volume of a metric ball in unitary group," *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6939 – 6949, 2015.

[6] R. Jin, T. Yang, M. Mahdavi, Y. F. Li, and Z. H. Zhou, "Improved bounds for the nystrm method with application to kernel classification," *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6939–6949, 2013.

[7] E. Yazdian, S. Gazor, M. H. Bastani, and M. Sharifitabar, "Eigenvalue estimation of the exponentially windowed sample covariance matrices," *IEEE Transactions on Information Theory*, vol. 62, no. 7, pp. 4300–4311, 2016.

[8] R. Couillet, F. Pascal, and J. W. Silverstein, "Robust estimates of covariance matrices in the large dimensional regime," *Information Theory IEEE Transactions on*, vol. 60, no. 11, pp. 7269–7278, 2012.

[9] H.-C. Cheng and M.-H. Hsieh, "Characterizations of matrix and operator-valued -entropies, and operator efron–stein inequalities," *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 472, no. 2187, 2016.

[10] H.-C. Cheng, M.-H. Hsieh, and M. Tomamichel, "Exponential decay of matrix -entropies on markov semigroups with applications to dynamical evolutions of quantum ensembles," *Journal of Mathematical Physics*, vol. 58, no. 9, p. 092202, 2017.

[11] H.-C. Cheng, M.-H. Hsieh, and P.-C. Yeh, "The learnability of unknown quantum measurements," *Quantum Info. Comput.*, vol. 16, no. 7-8, pp. 615–656, May 2016.

[12] J. A. Tropp, "An introduction to matrix concentration inequalities," *Foundations and Trends in Machine Learning*, vol. 8, no. 1-2, pp. 1–230, 2015.

[13] R. Ahlswede and A. Winter, "Strong converse for identification via quantum channels," *IEEE Transactions on Information Theory*, vol. 48, no. 3, pp. 569–579, 2002.

[14] D. Hsu, S. M. Kakade, and T. Zhang, "Tail inequalities for sums of random matrices that depend on the intrinsic dimension," *Electronic Communications in Probability*, vol. 17, no. 14, pp. 1–13, 2012.

[15] S. Minsker, "On some extensions of bernstein's inequality for self-adjoint operators," *Statistics & Probability Letters*, vol. 127, 2017.

[16] M. W. Meckes, "Concentration of norms and eigenvalues of random matrices," *Journal of Functional Analysis*, vol. 211, no. 2, pp. 508–524, 2004.

[17] L. Mackey, M. I. Jordan, R. Y. Chen, B. Farrell, J. A. Tropp *et al.*, "Matrix concentration inequalities via the method of exchangeable pairs," *Annals of Probability*, vol. 42, no. 3, pp. 906–945, 2014.

[18] D. Paulin, L. Mackey, and J. A. Tropp, "Deriving matrix concentration inequalities from kernel couplings," *arXiv preprint arXiv:1305.0612*, 2013.

[19] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *arXiv preprint arXiv:1011.3027*, 2010.

[20] M. Chiani, "On the probability that all eigenvalues of gaussian, wishart, and double wishart random matrices lie within an interval," *IEEE Transactions on Information Theory*, vol. 63, no. 7, pp. 4521–4531, 2017.

[21] L. Zhao, S. Liao, Y. Wang, Z. Li, J. Tang, and B. Yuan, "Theoretical properties for neural networks with weight matrices of low displacement rank," in *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[22] K. Choromanski and V. Sindhwani, "Recycling randomness with structure for sublinear time kernel expansions," in *Proceedings of the 33th International Conference on Machine Learning*, 2016, pp. 2502–2510.

[23] Y. Cheng, F. X. Yu, R. S. Feris, S. Kumar, A. Choudhary, and S.-F. Chang, "An exploration of parameter redundancy in deep networks with circulant projections," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2857–2865.

[24] A. Nemirovski, "Sums of random symmetric matrices and quadratic optimization under orthogonality constraints," *Mathematical Programming*, vol. 109, no. 2, pp. 283–317, 2007.

[25] A. M.-C. So, "Moment inequalities for sums of random matrices and their applications in optimization," *Mathematical Programming*, vol. 130, no. 1, pp. 125–151, 2011.

[26] A. Bose, A. Dasgupta, and H. Rubin, "A contemporary review and bibliography of infinitely divisible distributions and processes," *Indian Journal of Statistics, Series A*, vol. 64, no. 3, pp. 763–819, 2002.

[27] A. Kyprianou, *Introductory lectures on fluctuations of Lévy processes with applications*. Springer Science & Business Media, 2006.

[28] P. Chainais, "Multi-dimensional infinitely divisible cascades to model the statistics of natural images," in *IEEE International Conference on Image Processing 2005*, vol. 3. IEEE, 2005, pp. III–129.

[29] Y. Nishiyama and K. Fukumizu, "Characteristic kernels and infinitely divisible distributions," *Journal of Machine Learning Research*, vol. 17, no. 180, pp. 1–28, 2016.

[30] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[31] C. Houdré, "Remarks on deviation inequalities for functions of infinitely divisible random vectors," *Annals of Probability*, pp. 1223–1237, 2002.

[32] C. Zhang and D. Tao, "Generalization bound for infinitely divisible empirical process," in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, 2011, pp. 864–872.

[33] ——, "Risk bounds of learning processes for lévy processes," *Journal of Machine Learning Research*, vol. 14, no. Feb, pp. 351–376, 2013.

[34] C. Zhang, "Bennett-type generalization bounds: Large-deviation case and faster rate of convergence," in *Uncertainty in Artificial Intelligence*, 2013, p. 714.

[35] C. Houdré, V. Pérez-Abreu, and D. Surgailis, "Interpolation, correlation identities, and inequalities for infinitely divisible variables," *Journal of Fourier Analysis and Applications*, vol. 4, no. 6, pp. 651–668, 1998.

[36] K. Sato, *Lévy Processes and Infinitely Divisible Distributions*. Cambridge University Press, 1999.