Electron. Commun. Probab. **17** (2012), no. 14, 1–13.
DOI: 10.1214/ECP.v17-1869
ISSN: 1083-589X

ELECTRONIC COMMUNICATIONS in PROBABILITY

Tail inequalities for sums of random matrices that depend on the intrinsic dimension

Daniel Hsu* Sham M. Kakade† Tong Zhang‡

Abstract

This work provides exponential tail inequalities for sums of random matrices that depend only on intrinsic dimensions rather than explicit matrix dimensions. These tail inequalities are similar to the matrix versions of the Chernoff bound and Bernstein inequality except with the explicit matrix dimensions replaced by a trace quantity that can be small even when the explicit dimensions are large or infinite. Some applications to covariance estimation and approximate matrix multiplication are given to illustrate the utility of the new bounds.

Keywords: Large deviation; random matrix; intrinsic dimension.
AMS MSC 2010: Primary 60B20, Secondary 60F10.
Submitted to ECP on April 14, 2011, final version accepted on March 10, 2012.

1 Introduction

Sums of random matrices arise in many statistical and probabilistic applications, and hence their concentration behavior is of significant interest. Surprisingly, the classical exponential moment method used to derive tail inequalities for scalar random variables carries over to the matrix setting when augmented with certain matrix trace inequalities. This fact was first discovered by Ahlswede and Winter [1], who proved a matrix version of the Chernoff bound using the Golden-Thompson inequality [7, 23]: $\operatorname{tr}\exp(A+B) \leq \operatorname{tr}(\exp(A)\exp(B))$ for all symmetric matrices A and B. Later, it was demonstrated that the same technique could be adapted to yield analogues of other tail bounds such as Azuma's and Bernstein's inequalities [3, 9, 19, 8, 16, 15]. Recently, a theorem due to Lieb [12] was identified by Tropp [25, 24] to yield sharper versions of this general class of tail bounds. Altogether, these results have proved invaluable in constructing and simplifying many probabilistic arguments concerning sums of random matrices.

One deficiency of many of these previous inequalities is their dependence on the explicit matrix dimension (this is reviewed in Section 3.3), which prevents their application to infinite dimensional spaces that arise in a variety of data analysis tasks (*e.g.*, [22, 18, 6, 2, 11]). In this work, we prove analogous results where the dimension is replaced with a trace quantity that can be small even when the explicit matrix

*Microsoft Research New England.
E-mail: dahsu@microsoft.com
†Microsoft Research New England; and Department of Statistics, University of Pennsylvania.
E-mail: skakade@microsoft.com
‡Department of Statistics, Rutgers University.
E-mail: tzhang@stat.rutgers.edu

dimension is large or infinite. For instance, in our matrix generalization of Bernstein's inequality, the (scaled) trace of the second moment matrix appears instead of the matrix dimension. Such trace quantities can often be regarded as notions of intrinsic dimension. The price for this improvement is that the more typical exponential tail e^{-t} for $t > 0$ is replaced with a slightly weaker tail $t(e^t - t - 1)^{-1} \approx e^{-t+\log t}$. As t becomes large, the difference becomes negligible. For instance, if $t \geq 2.6$, then $t(e^t - t - 1)^{-1} \leq e^{-t/2}$.

There are some previous works that also give tail inequalities for sums of random matrices that do not depend on the explicit matrix dimension, at least in some special cases. For instance, in the case where the summands all have rank one, then Oliveira [16] gives a bound where the dimension is replaced by the number of summands. Rudelson and Vershynin [21] also prove similar exponential tail inequalities for sums of rank-one matrices using non-commutative Khinchine moment inequalities by way of a key inequality of Rudelson [20]; the extension to higher rank random matrices is not explicitly worked out, but may be possible. Indeed, Magen and Zouzias [14] pursue this direction, but their argument is complicated and falls short of giving an exponential tail inequality—this point is discussed in Section 3.3.

To concretely compare the new technique to previous results based on the matrix exponential moment method, consider the sum $\sum_{i=1}^n \gamma_i A_i$, where $A_1, A_2, \ldots, A_n$ are fixed symmetric $d \times d$ matrices, and $\gamma_1, \gamma_2, \ldots, \gamma_n$ are independent standard normal random variables. Tropp gives the following tail bound for the largest eigenvalue of the sum (Theorem 4.1 in [25]):

$$\Pr\left[\lambda_{\max}\left(\sum_{i=1}^n \gamma_i A_i\right) \geq \sqrt{2\|\Sigma\|_2 t}\right] \leq d \cdot e^{-t}$$

where $\Sigma := \sum_{i=1}^n A_i^2$. Combining Theorem 3.2 with Lemma 4.3 in [25] gives the following new tail bound:

$$\Pr\left[\lambda_{\max}\left(\sum_{i=1}^n \gamma_i A_i\right) \geq \sqrt{2\|\Sigma\|_2 t}\right] \leq \frac{\operatorname{tr}(\Sigma)}{\lambda_{\max}(\Sigma)} \cdot \frac{t}{e^t - t - 1}.$$

Note that $\operatorname{tr}(\Sigma)/\lambda_{\max}(\Sigma) \leq d$ but $t(e^t - t - 1)^{-1} > e^{-t}$ for $t > 0$, so no bound always dominates the other. However, for moderately large values of t, and when $d \gg \operatorname{tr}(\Sigma)/\lambda_{\max}(\Sigma)$, the new bound is a significant improvement.

2 Preliminaries

Let $\xi_1, \xi_2, \ldots, \xi_n$ be random variables, and for each $i = 1, 2, \ldots, n$, let

$$X_i := X_i(\xi_1, \xi_2, \ldots, \xi_i)$$

be a symmetric matrix-valued functional of $\xi_1, \xi_2, \ldots, \xi_i$. Assume the X_i have the same common range. We use $\mathbb{E}_i[\,\cdot\,]$ as shorthand for $\mathbb{E}[\,\cdot \mid \xi_1, \xi_2, \ldots, \xi_{i-1}]$. For any symmetric matrix H, let $\lambda_{\max}(H)$ denote its largest eigenvalue, $\exp(H) := I + \sum_{k=1}^\infty H^k/k!$, and $\log(\exp(H)) := H$.

The following convex trace inequality of Lieb [12] was also used by Tropp [25, 24].

Theorem 2.1 ([12])**.** *For any symmetric matrix H, the function $M \mapsto \operatorname{tr}\exp(H + \log(M))$ is concave in M for $M \succ 0$.*

The following lemma due to Tropp [24] is a matrix generalization of a scalar result due to Freedman [5] (see also [28]), where the key is the invocation of Theorem 2.1. We give the proof for completeness.

Lemma 2.2 ([24]). *Let I be the identity matrix for the range of the X_i. Then*

$$\mathbb{E}\left[\operatorname{tr}\left(\exp\left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \ln \mathbb{E}_i\left[\exp(X_i)\right]\right) - I\right)\right] \leq 0. \tag{2.1}$$

Proof. The proof is by induction on n. The claim holds trivially for $n = 0$. Now fix $n \geq 1$, and assume as the inductive hypothesis that (2.1) holds with n replaced by $n-1$. In this case,

$$\begin{aligned}
&\mathbb{E}\left[\operatorname{tr}\left(\exp\left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \log \mathbb{E}_i\left[\exp(X_i)\right]\right) - I\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}_n\left[\operatorname{tr}\left(\exp\left(\sum_{i=1}^{n-1} X_i - \sum_{i=1}^{n} \log \mathbb{E}_i\left[\exp(X_i)\right] + \log\exp(X_n)\right) - I\right)\right]\right] \\
&\leq \mathbb{E}\left[\operatorname{tr}\left(\exp\left(\sum_{i=1}^{n-1} X_i - \sum_{i=1}^{n} \log \mathbb{E}_i\left[\exp(X_i)\right] + \log \mathbb{E}_n\left[\exp(X_n)\right]\right) - I\right)\right] \\
&= \mathbb{E}\left[\operatorname{tr}\left(\exp\left(\sum_{i=1}^{n-1} X_i - \sum_{i=1}^{n-1} \log \mathbb{E}_i\left[\exp(X_i)\right]\right) - I\right)\right] \\
&\leq 0
\end{aligned}$$

where the first inequality follows from Theorem 2.1 and Jensen's inequality, and the second inequality follows from the inductive hypothesis. □

3 Exponential tail inequalities for sums of random matrices

3.1 A generic inequality

We first state a generic inequality based on Lemma 2.2. This differs from earlier approaches, which instead combine Markov's inequality with a result similar to Lemma 2.2 (c.f. Theorem 3.6 in [25]).

Theorem 3.1. *For any $\eta \in \mathbb{R}$ and any $t > 0$,*

$$\begin{aligned}
&\Pr\left[\lambda_{\max}\left(\eta \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \log \mathbb{E}_i\left[\exp(\eta X_i)\right]\right) > t\right] \\
&\qquad\qquad \leq \operatorname{tr}\left(\mathbb{E}\left[-\eta \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} \log \mathbb{E}_i\left[\exp(\eta X_i)\right]\right]\right) \cdot (e^t - t - 1)^{-1}.
\end{aligned}$$

Proof. Let $A := \eta \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \log \mathbb{E}_i[\exp(\eta X_i)]$. Note that $g(x) := e^x - x - 1$ is nonnegative for all $x \in \mathbb{R}$ and increasing for $x \geq 0$. Letting $\{\lambda_i(A)\}$ denote the eigenvalues of A, we have

$$\begin{aligned}
\Pr\left[\lambda_{\max}(A) > t\right](e^t - t - 1) &= \mathbb{E}\left[\mathbb{1}_{\{\lambda_{\max}(A) > t\}} \cdot (e^t - t - 1)\right] \\
&\leq \mathbb{E}\left[e^{\lambda_{\max}(A)} - \lambda_{\max}(A) - 1\right] \\
&\leq \mathbb{E}\left[\sum_i \left(e^{\lambda_i(A)} - \lambda_i(A) - 1\right)\right] \\
&= \mathbb{E}\left[\operatorname{tr}(\exp(A) - A - I)\right] \\
&\leq \operatorname{tr}(\mathbb{E}[-A])
\end{aligned}$$

where the last inequality $\mathbb{E}[\operatorname{tr}(\exp(A) - I)] \leq 0$ follows from Lemma 2.2. □

When $\sum_{i=1}^n X_i$ has zero-mean, then the first sum in the right-hand side of the inequality from Theorem 3.1 vanishes, so the trace is only over a sum of matrix logarithmic moment generating functions

$$\operatorname{tr}\left(\mathbb{E}\left[\sum_{i=1}^n \log \mathbb{E}_i\left[\exp(\eta X_i)\right]\right]\right).$$

For an appropriate choice of η, this trace quantity can be small even when the X_i have large or infinite explicit dimension.

3.2 Some specific bounds

We now give some specific bounds as corollaries of Theorem 3.1. The proofs use Theorem 3.1 together with some techniques from previous works (*e.g.*, [1, 25]) to yield new tail inequalities that depend on intrinsic notions of dimension rather than the explicit matrix dimensions.

First, we give a bound under a subgaussian-type condition on the distribution.

Theorem 3.2 (Matrix subgaussian bound)**.** *If there exists $\bar{\sigma} > 0$ and $\bar{k} > 0$ such that for all $i = 1, \dots, n$,*

$$\mathbb{E}_i[X_i] = 0, \quad \lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n \log \mathbb{E}_i\left[\exp(\eta X_i)\right]\right) \le \frac{\eta^2\bar{\sigma}^2}{2}, \quad \textit{and}$$
$$\mathbb{E}\left[\operatorname{tr}\left(\frac{1}{n}\sum_{i=1}^n \log \mathbb{E}_i\left[\exp(\eta X_i)\right]\right)\right] \le \frac{\eta^2\bar{\sigma}^2\bar{k}}{2}$$

for all $\eta > 0$ almost surely; then for any $t > 0$,

$$\Pr\left[\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) > \sqrt{\frac{2\bar{\sigma}^2 t}{n}}\right] \le \bar{k} \cdot t(e^t - t - 1)^{-1}.$$

Proof. We fix $\eta := \sqrt{2t/(\bar{\sigma}^2 n)}$. By Theorem 3.1, we obtain

$$\begin{aligned}
&\Pr\left[\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{n\eta}\sum_{i=1}^n \log \mathbb{E}_i\left[\exp(\eta X_i)\right]\right) > \frac{t}{n\eta}\right] \\
&\le \operatorname{tr}\left(\mathbb{E}\left[\sum_{i=1}^n \log \mathbb{E}_i\left[\exp(\eta X_i)\right]\right]\right) \cdot (e^t - t - 1)^{-1} \\
&\le \frac{n\eta^2\bar{\sigma}^2\bar{k}}{2} \cdot (e^t - t - 1)^{-1} \\
&= \bar{k} \cdot t(e^t - t - 1)^{-1}.
\end{aligned}$$

Now suppose

$$\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{n\eta}\sum_{i=1}^n \log \mathbb{E}_i\left[\exp(\eta X_i)\right]\right) \le \frac{t}{n\eta}.$$

By the sub-additivity of the map $M \mapsto \lambda_{\max}(M)$—*i.e.*, $\lambda_{\max}(A) \le \lambda_{\max}(B) + \lambda_{\max}(A - B)$—it follows that

$$\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) \le \lambda_{\max}\left(\frac{1}{n\eta}\sum_{i=1}^n \log \mathbb{E}_i\left[\exp(\eta X_i)\right]\right) + \frac{t}{n\eta} \le \frac{\eta\bar{\sigma}^2}{2} + \frac{t}{n\eta} = \sqrt{\frac{2\bar{\sigma}^2 t}{n}}. \quad \square$$

We can also give a Bernstein-type bound based on moment conditions. For simplicity, we just state the bound in the case that the $\lambda_{\max}(X_i)$ are bounded almost surely.

Theorem 3.3 (Matrix Bernstein bound)**.** *If there exists $\bar{b} > 0$, $\bar{\sigma} > 0$, and $\bar{k} > 0$ such that for all $i = 1, \dots, n$,*

$$\mathbb{E}_i[X_i] = 0, \quad \lambda_{\max}(X_i) \leq \bar{b}$$
$$\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_i[X_i^2]\right) \leq \bar{\sigma}^2, \quad \textit{and} \quad \mathbb{E}\left[\operatorname{tr}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_i[X_i^2]\right)\right] \leq \bar{\sigma}^2\bar{k}$$

almost surely; then for any $t > 0$,

$$\Pr\left[\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) > \sqrt{\frac{2\bar{\sigma}^2 t}{n}} + \frac{\bar{b}t}{3n}\right] \leq \bar{k} \cdot t(e^t - t - 1)^{-1}.$$

Proof. Let $\eta > 0$. For each $i = 1, \dots, n$,

$$\exp(\eta X_i) \preceq I + \eta X_i + \frac{e^{\eta\bar{b}} - \eta\bar{b} - 1}{\bar{b}^2} \cdot X_i^2$$

and therefore, by the operator monotonicity of the matrix logarithm and the fact $\log(1 + x) \leq x$,

$$\log \mathbb{E}_i\big[\exp(\eta X_i)\big] \preceq \frac{e^{\eta\bar{b}} - \eta\bar{b} - 1}{\bar{b}^2} \cdot \mathbb{E}_i\big[X_i^2\big].$$

Since $e^x - x - 1 \leq x^2/(2(1 - x/3))$ for $0 \leq x < 3$, we have by Theorem 3.1 and the subadditivity of the map $M \mapsto \lambda_{\max}(M)$,

$$\Pr\left[\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) > \frac{\eta\bar{\sigma}^2}{2(1 - \eta\bar{b}/3)} + \frac{t}{\eta n}\right] \leq \frac{\eta^2\bar{\sigma}^2\bar{k}n}{2(1 - \eta\bar{b}/3)} \cdot (e^t - t - 1)^{-1}$$

provided that $\eta < 3/\bar{b}$. Choosing

$$\eta := \frac{3}{\bar{b}} \cdot \left(1 - \frac{\sqrt{2\bar{\sigma}^2 t/n}}{2\bar{b}t/(3n) + \sqrt{2\bar{\sigma}^2 t/n}}\right)$$

gives the desired bound. □

3.3 Discussion

The results of this paper can be viewed as another sharpening of the matrix exponential moment method for deriving exponential tail inequalities for sums of random matrices, which has its origins in the work of Ahlswede and Winter [1] and was subsequently generalized and improved by many others [3, 9, 19, 8, 16, 15, 25, 24]. The novel feature of our results when compared to previous results is the absence of explicit dependence on the matrix dimensions. Indeed, nearly all previous tail inequalities using the exponential moment method (either via the Golden-Thompson inequality or Lieb's trace inequality) are roughly of the form $d \cdot e^{-t}$ when the matrices in the sum are $d \times d$ [1, 3, 9, 19, 8, 15, 25, 24]. For instance, a corollary of the "Master Tail Bound for Independent Sums" of Tropp (Theorem 3.6 in [25]) can be written as

$$\Pr\left[\lambda_{\max}\left(\eta\sum_{i=1}^{n} X_i\right) > \lambda_{\max}\left(\sum_{i=1}^{n} \log \mathbb{E}[\exp(\eta X_i)]\right) + t\right] \leq d \cdot e^{-t}$$

for all $t > 0$ and $\eta > 0$ (see Corollary 3.7 in [25]). Of course, when the random matrices are always confined to a single lower-dimensional space, then these previous results clearly depend only on this lower dimension (*i.e.*, d is replaced by this lower dimension). However, this situation is significantly less general than what is required in many

applications that involve very high or infinite dimensional matrices, such as the analysis of ridge regression [11], kernel methods [22, 6, 2], and Gaussian process methods [18]. Our results therefore widen the applicability of the matrix exponential moment method to handle these cases.

Relative to the tail inequalities of Rudelson and Vershynin [21] and Oliveira [16], we note that our inequalities apply to random matrices of any rank, rather than just rank-one (or low-rank) random matrices. Although [21] and [16] only explicitly provides inequalities for the rank-one case, the work of Magen and Zouzias [14] gives an extension that applies to higher rank random matrices. [14] considers the specific case where $X_1, \ldots, X_n$ are i.i.d. copies of a random matrix X which satisfies $\|\mathbb{E}[X]\|_2 \leq 1$, $\|X\|_2 \leq \gamma$, and $\mathrm{rank}(X) \leq r$ almost surely; their bound has the following form:

$$\Pr\left[\left\|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X]\right\|_2 > \sqrt{\frac{c_2\gamma t \log n}{n}}\right] \leq r \cdot n \cdot \exp\left(-(c_1 \log n) \cdot (\log t)\right)$$

for some unspecified positive constants c_1 and c_2. It should be noted, however, that the right-hand side decreases only polynomially in t rather than exponentially, which is qualitatively weaker than the previous results of [21] and [16]; therefore, it is not a strict improvement or generalization.

One disadvantage of our technique is that in finite dimensional settings, the relevant trace quantity that replaces the dimension may turn out to be of the same order as the dimension d (an example of such a case is discussed next). In such cases, the resulting tail bound from Theorem 3.3 (say) of $\bar{k} \cdot t(e^t - t - 1)^{-1}$ is looser than the $d \cdot e^{-t}$ tail bound provided by earlier techniques [25], and this can be significant for small values of t.

We note that the general matrix exponential moment method used here and in previous work leads to a significantly suboptimal tail inequality in some cases. This was pointed out in [25], but we elaborate on it here further. Suppose $x_1, \ldots, x_n \in \{\pm 1\}^d$ are i.i.d. random vectors with independent Rademacher entries: each coordinate of x_i is $+1$ or -1 with equal probability. Let $X_i = x_i x_i^\top - I$, so $\mathbb{E}[X_i] = 0$, $\lambda_{\max}(X_i) = \lambda_{\max}(\mathbb{E}[X_i^2]) = d - 1$, and $\mathrm{tr}(\mathbb{E}[X_i^2]) = d(d-1)$. In this case, Theorem 3.3 implies the bound

$$\Pr\left[\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top - I\right) > \sqrt{\frac{2(d-1)t}{n}} + \frac{(d-1)t}{3n}\right] \leq d \cdot t(e^t - t - 1)^{-1}.$$

On the other hand, because the x_i have subgaussian projections, it is known that

$$\Pr\left[\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top - I\right) > 2\sqrt{\frac{71d + 16t}{n}} + \frac{10d + 2t}{n}\right] \leq 2e^{-t/2}$$

(see Lemma A.1 in Appendix A). First, this latter inequality removes the d factor on the right-hand side. But more importantly, the deviation term t does not scale with d in this inequality, whereas it does in the former. Thus this latter bound provides a much stronger exponential tail: roughly put, $\Pr[\lambda_{\max}(\sum_{i=1}^n x_i x_i^\top/n - I) > c \cdot (\sqrt{d/n} + d/n) + \tau] \leq \exp(-\Omega(n \min(\tau, \tau^2)))$ for some constant $c > 0$ (note that the dimension d does not appear in the exponent); the probability bound from Theorem 3.3 is only of the form $\exp(-\Omega((n/d)\min(\tau, \tau^2)))$. The sub-optimality of Theorem 3.3 is shared by all other existing tail inequalities proved using this exponential moment method. The issue may be related to the asymptotic freeness of the $d \times n$ random matrix $[x_1|x_2|\cdots|x_n]$ [27, 10]—*i.e.*, that nearly all high-order moments of random matrices with independent entries vanish asymptotically—which is not exploited in the matrix exponential moment method. This means that the proof technique in the exponential moment method overestimates the contribution of high-order matrix moments that should have vanished. Formalizing this

discrepancy would help clarify the limits of this technique, but the task is beyond the scope of this paper. It is also worth mentioning that this phenomenon only appears to hold when the x_i have independent entries (and other similar cases). In cases with correlated entries, our bound is close to best possible in the worst case.

4 Examples

For a matrix M, let $\|M\|_F$ denote its Frobenius norm, and let $\|M\|_2$ denote its spectral norm. If M is symmetric, then $\|M\|_2 = \max\{\lambda_{\max}(M),\ -\lambda_{\min}(M)\}$, where $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ are, respectively, the largest and smallest eigenvalues of M.

4.1 Supremum of a random process

The first example embeds a random process in a diagonal matrix to show that Theorem 3.2 is tight in certain cases.

Example 4.1. *Let $(z_1, z_2, \dots)$ be (possibly dependent) mean-zero subgaussian random variables;* i.e.*, each $\mathbb{E}[z_i] = 0$, and there exists positive constants $\sigma_1, \sigma_2, \dots$ such that*

$$\mathbb{E}[\exp(\eta z_i)] \le \exp\left(\frac{\eta^2\sigma_i^2}{2}\right) \quad \forall \eta \in \mathbb{R}.$$

We further assume that $v := \sup_i\{\sigma_i^2\} < \infty$ and $k := \frac{1}{v}\sum_i \sigma_i^2 < \infty$. Also, for convenience, we assume $\log k \ge 1.3$ (to simplify the tail inequality).

Let $X = \operatorname{diag}(z_1, z_2, \dots)$ be the random diagonal matrix with the z_i on its diagonal. We have $\mathbb{E}[X] = 0$, and

$$\log \mathbb{E}[\exp(\eta X)] \preceq \operatorname{diag}\left(\frac{\eta^2\sigma_1^2}{2}, \frac{\eta^2\sigma_2^2}{2}, \dots\right)$$

by the operator monotonicity of the matrix logarithm, so

$$\lambda_{\max}\left(\log \mathbb{E}[\exp(\eta X)]\right) \le \frac{\eta^2 v}{2} \quad \text{and} \quad \operatorname{tr}\left(\log \mathbb{E}[\exp(\eta X)]\right) \le \frac{\eta^2 vk}{2}.$$

By Theorem 3.2, we have

$$\Pr\left[\lambda_{\max}(X) > \sqrt{2vt}\right] \le k \cdot t(e^t - t - 1)^{-1}.$$

Therefore, letting $t := 2(\tau + \log k) > 2.6$ for $\tau > 0$ and interpreting $\lambda_{\max}(X)$ as $\sup_i\{z_i\}$,

$$\Pr\left[\sup_i\{z_i\} > 2\sqrt{\sup_i\{\sigma_i^2\}\left(\log\frac{\sum_i \sigma_i^2}{\sup_i\{\sigma_i^2\}} + \tau\right)}\right] \le e^{-\tau}.$$

Suppose the $z_i \sim \mathcal{N}(0,1)$ are just N i.i.d. standard Gaussian random variables. Then the above inequality states that the largest of the z_i is $O(\log N + \tau)$ with probability at least $1 - e^{-\tau}$; this is known to be tight up to constants, so the $\log N$ term cannot generally be removed. This fact has been noted by previous works on matrix tail inequalities [25], which also use this example as an extreme case. We note, however, that these previous works are not directly applicable to the case of a countably infinite number of mean-zero Gaussian random variables $z_i \sim \mathcal{N}(0, \sigma_i^2)$ (or more generally, subgaussian random variables), whereas the above inequality can be applied as long as the sum of the σ_i^2 is finite.

4.2 Covariance estimation

Our next example uses Theorem 3.3 to give a spectral norm error bound for estimating the second moment matrix of a random vector from i.i.d. copies. This is relevant in the context of (kernel) principal component analysis of high (or infinite) dimensional data [22].

Example 4.2. *Let $x_1, x_2, \ldots, x_n$ be i.i.d. random vectors with $\Sigma := \mathbb{E}[x_i x_i^\top]$, $K := \mathbb{E}[x_i x_i^\top x_i x_i^\top]$, and $\|x_i\|_2 \leq \ell$ almost surely for some $\ell > 0$. Let $X_i := x_i x_i^\top - \Sigma$ and $\hat{\Sigma}_n := n^{-1}\sum_{i=1}^n x_i x_i^\top$. We have $\lambda_{\max}(X_i) \leq \ell^2 - \lambda_{\min}(\Sigma)$. Also, $\lambda_{\max}(n^{-1}\sum_{i=1}^n \mathbb{E}[X_i^2]) = \lambda_{\max}(K - \Sigma^2)$ and $\mathbb{E}[\mathrm{tr}(n^{-1}\sum_{i=1}^n \mathbb{E}[X_i^2])] = \mathrm{tr}(K - \Sigma^2)$. By Theorem 3.3,*

$$\Pr\left[\lambda_{\max}\big(\hat{\Sigma}_n - \Sigma\big) > \sqrt{\frac{2\lambda_{\max}(K-\Sigma^2)t}{n}} + \frac{(\ell^2 - \lambda_{\min}(\Sigma))t}{3n}\right] \leq \frac{\mathrm{tr}(K-\Sigma^2)}{\lambda_{\max}(K-\Sigma^2)} \cdot \frac{t}{e^t - t - 1}.$$

Since $\lambda_{\max}(-X_i) \leq \lambda_{\max}(\Sigma)$, we also have

$$\Pr\left[\lambda_{\max}\big(\Sigma - \hat{\Sigma}_n\big) > \sqrt{\frac{2\lambda_{\max}(K-\Sigma^2)t}{n}} + \frac{\lambda_{\max}(\Sigma)t}{3n}\right] \leq \frac{\mathrm{tr}(K-\Sigma^2)}{\lambda_{\max}(K-\Sigma^2)} \cdot \frac{t}{e^t - t - 1}.$$

Therefore

$$\Pr\left[\big\|\hat{\Sigma}_n - \Sigma\big\|_2 > \sqrt{\frac{2\lambda_{\max}(K-\Sigma^2)t}{n}} + \frac{\max\{\ell^2 - \lambda_{\min}(\Sigma),\ \lambda_{\max}(\Sigma)\}t}{3n}\right] \leq \frac{\mathrm{tr}(K-\Sigma^2)}{\lambda_{\max}(K-\Sigma^2)} \cdot \frac{2t}{e^t - t - 1}.$$

Above, the relevant notion of intrinsic dimension is $\mathrm{tr}(K-\Sigma^2)/\lambda_{\max}(K-\Sigma^2)$, which can be finite even when the random vectors x_i take on values in an infinite dimensional Hilbert space. A related result was given in [29] for Frobenius norm error $\|\hat{\Sigma}_n - \Sigma\|_F$ rather than spectral norm error. This is generally incomparable to our result, although spectral norm error may be more appropriate in cases where the spectrum is slow to decay.

4.3 Approximate matrix multiplication

Finally, we give an example about approximating a matrix product $AB^\top$ using non-uniform sampling of the columns of A and B.

Example 4.3. *Let $A := [a_1|a_2|\cdots|a_m]$ and $B := [b_1|b_2|\cdots|b_m]$ be fixed matrices, each with m columns. Assume $a_i \neq 0$ and $b_i \neq 0$ for all $i = 1, 2, \ldots, m$. If m is very large, then the straightforward computation of the product $AB^\top$ can be prohibitive. An alternative is to take a small (non-uniform) random sample of the columns of A and B, say $a_{i_1}, b_{i_1}, a_{i_2}, b_{i_2}, \ldots, a_{i_n}, b_{i_n}$, and then compute a weighted sum of outer products*

$$\frac{1}{n}\sum_{j=1}^n \frac{a_{i_j} b_{i_j}^\top}{p_{i_j}}$$

where $p_{i_j} > 0$ is the a priori *probability of choosing the column index $i_j \in \{1, 2, \ldots, m\}$ (the actual values of the probabilities p_i for $i = 1, 2, \ldots, m$ are given below). An analysis of this randomized approximation scheme is given below. This scheme was originally proposed and analyzed by Drineas, Kannan, and Mahoney [4], where the error measure used was the Frobenius norm; here, we analyze the spectral norm error. The spectral*

norm error was also analyzed in [14], but the result had a worse dependence on the allowed failure probability.

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random matrices with the discrete distribution given by

$$\Pr\left[X_j = \frac{1}{p_i}\begin{bmatrix} 0 & a_i b_i^\top \\ b_i a_i^\top & 0 \end{bmatrix}\right] = p_i \propto \|a_i\|_2\|b_i\|_2$$

for all $i = 1, 2, \ldots, m$, where $p_i := \|a_i\|_2\|b_i\|_2/Z$ and $Z := \sum_{i=1}^m \|a_i\|_2\|b_i\|_2$. Let

$$\hat{M}_n := \frac{1}{n}\sum_{j=1}^n X_j \quad \text{and} \quad M := \begin{bmatrix} 0 & AB^\top \\ BA^\top & 0 \end{bmatrix}.$$

Note that $\|\hat{M}_n - M\|_2$ is the spectral norm error of approximating $AB^\top$ using the average of n outer products $\sum_{j=1}^n a_{ij} b_{ij}^\top / p_{i_j}$, where the indices are such that $i_j = i \Leftrightarrow X_j = a_i b_i^\top / p_i$ for $j = 1, 2, \ldots, n$.

We have the following identities:

$$\begin{aligned}
\mathbb{E}[X_j] &= \sum_{i=1}^m p_i \left(\frac{1}{p_i}\begin{bmatrix} 0 & a_i b_i^\top \\ b_i a_i^\top & 0 \end{bmatrix}\right) = \begin{bmatrix} 0 & \sum_{i=1}^m a_i b_i^\top \\ \sum_{i=1}^m b_i a_i^\top & 0 \end{bmatrix} = M \\
\operatorname{tr}(\mathbb{E}[X_j^2]) &= \operatorname{tr}\left(\sum_{i=1}^m p_i \left(\frac{1}{p_i^2}\begin{bmatrix} a_i b_i^\top b_i a_i^\top & 0 \\ 0 & b_i a_i^\top a_i b_i^\top \end{bmatrix}\right)\right) = \sum_{i=1}^m \frac{2\|a_i\|_2^2\|b_i\|_2^2}{p_i} = 2Z^2 \\
\operatorname{tr}(\mathbb{E}[X_j]^2) &= \operatorname{tr}\left(\begin{bmatrix} AB^\top BA^\top & 0 \\ 0 & BA^\top AB^\top \end{bmatrix}\right) = 2\operatorname{tr}(A^\top AB^\top B);
\end{aligned}$$

and the following inequalities:

$$\begin{aligned}
\|X_j\|_2 &\le \max_{i=1,\ldots,m} \frac{1}{p_i}\left\|\begin{bmatrix} 0 & a_i b_i^\top \\ b_i a_i^\top & 0 \end{bmatrix}\right\|_2 = \max_{i=1,\ldots,m} \frac{\|a_i b_i^\top\|_2}{p_i} = Z \\
\|\mathbb{E}[X_j]\|_2 &= \|AB^\top\|_2 \le \|A\|_2\|B\|_2 \\
\|\mathbb{E}[X_j^2]\|_2 &\le \|A\|_2\|B\|_2 Z.
\end{aligned}$$

This means $\|X_j - M\|_2 \le Z + \|A\|_2\|B\|_2$ and $\|\mathbb{E}[(X_j - M)^2]\|_2 \le \|\mathbb{E}[X_j^2] - M^2\|_2 \le \|A\|_2\|B\|_2(Z + \|A\|_2\|B\|_2)$, so Theorem 3.3 and a union bound imply

$$\begin{aligned}
\Pr\left[\left\|\hat{M}_n - M\right\|_2 > \sqrt{\frac{2\left(\|A\|_2\|B\|_2(Z + \|A\|_2\|B\|_2)\right)t}{n}} + \frac{(Z + \|A\|_2\|B\|_2)t}{3n}\right] \\
\le 4\left(\frac{Z^2 - \operatorname{tr}(A^\top AB^\top B)}{\|A\|_2\|B\|_2(Z + \|A\|_2\|B\|_2)}\right)\cdot\frac{t}{e^t - t - 1}.
\end{aligned}$$

Let $r_A := \|A\|_F^2/\|A\|_2^2 \in [1, \operatorname{rank}(A)]$ and $r_B := \|B\|_F^2/\|B\|_2^2 \in [1, \operatorname{rank}(B)]$ be the numerical (or stable) rank of A and B, respectively. Since

$$\frac{Z}{\|A\|_2\|B\|_2} \le \frac{\|A\|_F\|B\|_F}{\|A\|_2\|B\|_2} = \sqrt{r_A r_B},$$

we have the simplified (but slightly looser) bound (for $t \ge 2.6$)

$$\Pr\left[\frac{\left\|\hat{M}_n - M\right\|_2}{\|A\|_2\|B\|_2} > \sqrt{\frac{2(1+\sqrt{r_A r_B})t}{n}} + \frac{(1+\sqrt{r_A r_B})t}{3n}\right] \le 4\sqrt{r_A r_B}\cdot e^{-t/2}.$$

Therefore, for any $\epsilon \in (0,1)$ and $\delta \in (0,1)$, if

$$n \ge \left(\frac{8}{3} + 2\sqrt{\frac{5}{3}}\right)\frac{(1+\sqrt{r_A r_B})\log(4\sqrt{r_A r_B}/\delta)}{\epsilon^2},$$

then with probability at least $1-\delta$ *over the random choice of column indices* $i_1, i_2, \ldots, i_n$*,*

$$\left\| \frac{1}{n} \sum_{j=1}^{n} \frac{a_{i_j} b_{i_j}^\top}{p_{i_j}} - AB^\top \right\|_2 \leq \epsilon \|A\|_2 \|B\|_2.$$

A Sums of random vector outer products

The following lemma is a tail inequality for the spectral norm error of the empirical covariance matrix of subgaussian random vectors. This result (without explicit constants) is due to Litvak, Pajor, Rudelson, and Tomczak-Jaegermann [13] (see also [26]).

Lemma A.1. *Let* $x_1, x_2, \ldots, x_n$ *be random vectors in* $\mathbb{R}^d$ *such that, for some* $\gamma \geq 0$*,*

$$\mathbb{E}\Big[x_i x_i^\top \;\Big|\; x_1, x_2, \ldots, x_{i-1} \Big] = I \quad \textit{and}$$

$$\mathbb{E}\Big[\exp\left(\alpha^\top x_i\right) \;\Big|\; x_1, x_2, \ldots, x_{i-1} \Big] \leq \exp\left(\|\alpha\|_2^2 \gamma / 2 \right), \quad \forall \alpha \in \mathbb{R}^d$$

for all $i = 1, 2, \ldots, n$*, almost surely. For all* $\epsilon_0 \in (0, 1/2)$ *and* $t > 0$*,*

$$\Pr\left[\left\| \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top - I \right\|_2 > \frac{1}{1 - 2\epsilon_0} \cdot \varepsilon_{\epsilon_0, t, n} \right] \leq 2e^{-t}$$

where

$$\varepsilon_{\epsilon_0, t, n} := \gamma \cdot \left(\sqrt{\frac{32\left(d \log(1 + 2/\epsilon_0) + t\right)}{n}} + \frac{2\left(d \log(1 + 2/\epsilon_0) + t\right)}{n} \right).$$

For completeness, we give a detailed proof of Lemma A.1 by applying the tail inequality in Lemma A.2 to Rayleigh quotients of the empirical covariance matrix, together with a covering argument based on the estimate in Lemma A.3 from [17].

Lemma A.2. *Let* $\xi_1, \xi_2, \ldots, \xi_n$ *be independent random variables such that*

$$\mathbb{E}_i[\exp(\eta \xi_i)] \leq \exp(\gamma \eta^2 / 2), \quad \forall \eta \in \mathbb{R}$$

for all $i = 1, 2, \ldots, n$*, almost surely. For any* $t > 0$*,*

$$\Pr\left[\frac{1}{n} \sum_{i=1}^{n} (\xi_i^2 - \mathbb{E}_i[\xi_i^2]) > \gamma \sqrt{\frac{32t}{n}} + \frac{2\gamma t}{n} \right] \leq e^{-t} \textit{ and } \Pr\left[\frac{1}{n} \sum_{i=1}^{n} (\xi_i^2 - \mathbb{E}_i[\xi_i^2]) < \gamma \sqrt{\frac{32t}{n}} \right] \leq e^{-t}.$$

Proof. Observe that $\mathbb{E}_i[\exp(\eta|\xi_i|)] \leq \mathbb{E}_i[\exp(\eta\xi_i) + \exp(-\eta\xi_i)] \leq 2\exp(\gamma\eta^2/2)$ for all $\eta \in \mathbb{R}$. By Chernoff's bounding method, $\mathbb{E}_i[\mathbb{1}_{\{\xi_i^2 > \varepsilon\}}] \leq 2e^{-\varepsilon/(2\gamma)}$ for all $\varepsilon \geq 0$. Therefore, for all $\eta < 1/(2\gamma)$,

$$\begin{aligned}
\mathbb{E}_i[\exp(\eta \xi_i^2)] &= 1 + \eta \mathbb{E}_i[\xi_i^2] + \eta \int_0^\infty (\exp(\eta\varepsilon) - 1) \mathbb{E}_i[\mathbb{1}_{\{\xi_i^2 > \varepsilon\}}] d\varepsilon \\
&\leq 1 + \eta \mathbb{E}_i[\xi_i^2] + 2\eta \int_0^\infty (\exp(\eta\varepsilon) - 1) \exp(-\varepsilon/(2\gamma)) d\varepsilon \\
&= 1 + \eta \mathbb{E}_i[\xi_i^2] + \frac{8\gamma^2\eta^2}{1 - 2\gamma\eta} \leq \exp\left(\eta \mathbb{E}_i[\xi_i^2] + \frac{8\gamma^2\eta^2}{1 - 2\gamma\eta} \right)
\end{aligned}$$

where the first equation follows from integration-by-parts. Since the above holds almost surely for all $i = 1, 2, \ldots, n$,

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\eta \sum_{i=1}^{n} (\xi_i^2 - \mathbb{E}_i[\xi_i^2]) \right) \right] &= \mathbb{E}\left[\exp\left(\eta \sum_{i=1}^{n-1} (\xi_i^2 - \mathbb{E}_i[\xi_i^2]) \right) \mathbb{E}_n\left[\exp\left(\eta(\xi_n^2 - \mathbb{E}_n[\xi_n^2]) \right) \right] \right] \\
&\leq \mathbb{E}\left[\exp\left(\eta \sum_{i=1}^{n-1} (\xi_i^2 - \mathbb{E}_i[\xi_i^2]) \right) \exp\left(\frac{8\gamma^2\eta^2}{1 - 2\gamma\eta} \right) \right] \leq \cdots \leq \exp\left(\frac{8\gamma^2 n \eta}{1 - 2\gamma\eta} \right).
\end{aligned}$$

By Chernoff's bounding method, for all $0 \leq \eta < 1/(2\gamma)$ and $\varepsilon \geq 0$,

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n}(\xi_i^2 - \mathbb{E}_i[\xi_i^2]) > \varepsilon\right] \leq \exp\left(-\eta n\varepsilon + \frac{8\gamma^2 n\eta^2}{1-2\gamma\eta}\right).$$

Setting $\eta := \frac{1}{2\gamma}\left(1-\sqrt{\frac{4\gamma}{4\gamma+\varepsilon}}\right)$ and $\varepsilon := \gamma\sqrt{32t/n}+2\gamma t/n$ gives the first claimed probability bound. Similarly, for all $\eta \leq 0$ and $\varepsilon \geq 0$,

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n}(\xi_i^2 - \mathbb{E}[\xi_i^2]) < -\varepsilon\right] \leq \exp\left(\eta n\varepsilon + \frac{8\gamma^2 n\eta^2}{1-2\gamma\eta}\right) \leq \exp\left(\eta n\varepsilon + 8\gamma^2 n\eta^2\right).$$

Setting $\eta := -\frac{\varepsilon}{16\gamma^2}$ and $\varepsilon := \gamma\sqrt{32t/n}$ gives the second claimed probability bound. □

Lemma A.3 ([17]). *For any $\epsilon_0 > 0$, there exists $Q \subseteq \mathcal{S}^{d-1} := \{\alpha \in \mathbb{R}^d : \|\alpha\|_2 = 1\}$ of cardinality $\leq (1+2/\epsilon_0)^d$ such that $\forall \alpha \in \mathcal{S}^{d-1} \exists q \in Q \,.\, \|\alpha - q\|_2 \leq \epsilon_0$.*

Proof of Lemma A.1. Let $\hat{\Sigma} := (1/n)\sum_{i=1}^{n} x_i x_i^\top$, let $\mathcal{S}^{d-1} := \{\alpha \in \mathbb{R}^d : \|\alpha\|_2 = 1\}$ be the unit sphere in $\mathbb{R}^d$, and let $Q \subset \mathcal{S}^{d-1}$ be an ϵ_0-cover of $\mathcal{S}^{d-1}$ with respect to $\|\cdot\|_2$ of minimum size. By Lemma A.3, the cardinality of Q is at most $(1+2/\epsilon_0)^d$.

Let E be the event in which $\max\left\{|q^\top(\hat{\Sigma}-I)q| : q \in Q\right\} \leq \varepsilon_{\epsilon_0,t,n}$. Observe that for all $q \in Q$, $\mathbb{E}[(q^\top x_i)^2 \mid x_1, x_2, \ldots, x_{i-1}] = 1 = q^\top q$, and $\mathbb{E}[\exp(\eta q^\top x_i)|x_1, x_2, \ldots, x_{i-1}] \leq \exp(\gamma\eta^2/2)$ for all $\eta \in \mathbb{R}$ and $i = 1, 2, \ldots, n$, almost surely. Therefore, by Lemma A.2 and a union bound, $\Pr[E] \geq 1 - 2e^{-t}$. Now assume the event E holds. Let $\alpha_0 \in \mathcal{S}^{d-1}$ be such that $|\alpha_0^\top(\hat{\Sigma}-I)\alpha_0| = \max\{|\alpha^\top(\hat{\Sigma}-I)\alpha| : \alpha \in \mathcal{S}^{d-1}\} = \|\hat{\Sigma}-I\|_2$. Using the triangle and Cauchy-Schwarz inequalities, we have

$$\begin{aligned}
\|\hat{\Sigma}-I\|_2 &= |\alpha_0^\top(\hat{\Sigma}-I)\alpha_0| \\
&\leq \min_{q\in Q} |q^\top(\hat{\Sigma}-I)q| + |\alpha_0^\top(\hat{\Sigma}-I)(\alpha_0-q)| + |(q-\alpha_0)^\top(\hat{\Sigma}-I)q| \\
&\leq \min_{q\in Q} |q^\top(\hat{\Sigma}-I)q| + \|\alpha_0\|_2\|\hat{\Sigma}-I\|_2\|\alpha_0-q\|_2 + \|q-\alpha_0\|_2\|\hat{\Sigma}-I\|_2\|q\|_2 \\
&\leq \varepsilon_{\epsilon_0,t,n} + 2\epsilon_0\|\hat{\Sigma}-I\|_2
\end{aligned}$$

so $\|\hat{\Sigma}-I\|_2 \leq \frac{1}{1-2\epsilon_0}\cdot\varepsilon_{\epsilon_0,t,n}$. □

References

[1] Rudolf Ahlswede and Andreas Winter, *Strong converse for identification via quantum channels*, IEEE Trans. Inform. Theory **48** (2002), no. 3, 569–579. MR-1889969

[2] Francis R. Bach, *Consistency of the group lasso and multiple kernel learning*, J. Mach. Learn. Res. **9** (2008), 1179–1225. MR-2417268

[3] Demetres Christofides and Klas Markström, *Expansion properties of random Cayley graphs and vertex transitive graphs via matrix martingales*, Random Structures Algorithms **32** (2008), no. 1, 88–100. MR-2371053

[4] Petros Drineas, Ravi Kannan, and Michael W. Mahoney, *Fast Monte Carlo algorithms for matrices. I. Approximating matrix multiplication*, SIAM J. Comput. **36** (2006), no. 1, 132–157. MR-2231643

[5] David A. Freedman, *On tail probabilities for martingales*, Ann. Probability **3** (1975), 100–118. MR-0380971

[6] Kenji Fukumizu, Francis R. Bach, and Arthur Gretton, *Statistical consistency of kernel canonical correlation analysis*, J. Mach. Learn. Res. **8** (2007), 361–383 (electronic). MR-2320675

[7] Sidney Golden, *Lower bounds for the Helmholtz function*, Phys. Rev. (2) **137** (1965), B1127–B1128. MR-0189691

[8] David Gross, *Recovering low-rank matrices from few coefficients in any basis*, IEEE Trans. Inform. Theory **57** (2011), no. 3, 1548–1566. MR-2815834

[9] David Gross, Yi-Kai Liu, Steven T. Flammia, Stephen Becker, and Jens Eisert, *Quantum state tomography via compressed sensing*, Phys. Rev. Lett. **105** (2010), 150401.

[10] Alice Guionnet, *Large deviations and stochastic calculus for large random matrices*, Probab. Surv. **1** (2004), 72–172. MR-2095566

[11] Daniel Hsu, Sham M. Kakade, and Tong Zhang, *An analysis of random design linear regression*, 2011, arXiv:1106.2363v1.

[12] Elliott H. Lieb, *Convex trace functions and the Wigner-Yanase-Dyson conjecture*, Advances in Math. **11** (1973), 267–288. MR-0332080

[13] Alexander E. Litvak, Alain Pajor, Mark Rudelson, and Nicole Tomczak-Jaegermann, *Smallest singular value of random matrices and geometry of random polytopes*, Adv. Math. **195** (2005), no. 2, 491–523. MR-2146352

[14] Avner Magen and Anastasios Zouzias, *Low rank matrix-valued Chernoff bounds and approximate matrix multiplication*, Proceedings of the 22nd ACM-SIAM Symposium on Discrete Algorithms, 2011.

[15] Roberto Imbuzeiro Oliveira, *Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges*, 2010, arXiv:0911.0600.

[16] ———, *Sums of random Hermitian matrices and an inequality by Rudelson*, Electron. Commun. Probab. **15** (2010), 203–212. MR-2653725

[17] Gilles Pisier, *The volume of convex bodies and Banach space geometry*, Cambridge Tracts in Mathematics, vol. 94, Cambridge University Press, Cambridge, 1989. MR-1036275

[18] Carl Edward Rasmussen and Christopher K. I. Williams, *Gaussian processes for machine learning*, Adaptive Computation and Machine Learning, MIT Press, Cambridge, MA, 2006. MR-2514435

[19] Benjamin Recht, *A simpler approach to matrix completion*, J. Mach. Learn. Res. **12** (2011), 3413–3430.

[20] Mark Rudelson, *Random vectors in the isotropic position*, J. Funct. Anal. **164** (1999), no. 1, 60–72. MR-1694526

[21] Mark Rudelson and Roman Vershynin, *Sampling from large matrices: an approach through geometric functional analysis*, J. ACM **54** (2007), no. 4, Art. 21, 19 pp. (electronic). MR-2351844

[22] Bernhard Schölkopf, Alex J. Smola, and Klaus-Robert Müller, *Kernel principal component analysis*, Advances in Kernel Methods—Support Vector Learning (Bernhard Schölkopf, Christopher J. C. Burges, and Alex J. Smola, eds.), MIT Press, 1999, pp. 327–352.

[23] Colin J. Thompson, *Inequality with applications in statistical mechanics*, J. Mathematical Phys. **6** (1965), 1812–1813. MR-0189688

[24] Joel A. Tropp, *Freedman's inequality for matrix martingales*, Electron. Commun. Probab. **16** (2011), 262–270. MR-2802042

[25] ———, *User-friendly tail bounds for sums of random matrices*, Foundations of Computational Mathematics (2011), 1–46.

[26] Roman Vershynin, *Introduction to the non-asymptotic analysis of random matrices*, Compressed Sensing, Theory and Applications (Y. Eldar and G. Kutyniok, eds.), Cambridge University Press, 2012, pp. 210–268.

[27] Dan Voiculescu, *Limit laws for random matrices and free products*, Invent. Math. **104** (1991), no. 1, 201–220. MR-1094052

[28] Tong Zhang, *Data dependent concentration bounds for sequential prediction algorithms*, Learning theory, Lecture Notes in Comput. Sci., vol. 3559, Springer, Berlin, 2005, pp. 173–187. MR-2203261

[29] Laurent Zwald, Olivier Bousquet, and Gilles Blanchard, *Statistical properties of kernel principal component analysis*, Learning theory, Lecture Notes in Comput. Sci., vol. 3120, Springer, Berlin, 2004, pp. 594–608. MR-2177937

Acknowledgments. We are grateful to Alex Gittens, Joel Tropp, and the anonymous reviewer for their many comments and suggestions.