Bernoulli **12**(5), 2006, 761–773

On singular values of matrices with independent rows

SHAHAR MENDELSON[1] and ALAIN PAJOR[2]

[1]*Centre for Mathematics and its Applications, Institute of Advanced Studies, Australian National University, Canberra, ACT 0200, Australia. E-mail: shahar.mendelson@anu.edu.au*
[2]*Equipe d'Analyse et Mathématiques Appliquées, Université de Marne-la-Vallée, 5 boulevard Descartes, Champs sur Marne, 77454 Marne-la-Vallée Cedex 2, France.*
E-mail: alain.pajor@univ-mlv.fr

We present deviation inequalities of random operators of the form $N^{-1}\sum_{i=1}^{N} X_i \otimes X_i$ from the average operator $\mathbb{E}(X \otimes X)$, where X_i are independent random vectors distributed as X, which is a random vector in $\mathbb{R}^n$ or in ℓ_2. We use these inequalities to estimate the singular values of random matrices with independent rows (without assuming that the entries are independent).

Keywords: random vectors in $\mathbb{R}^n$; singular values of integral operators

1. Introduction

The goal of this paper is to present deviation inequalities of random operators defined via independent copies of a random vector X in $\mathbb{R}^n$ or in the Hilbert space ℓ_2. To be more exact (and for the sake of simplicity), let X be a random vector taking values in $\mathbb{R}^n$ and consider $(X_i)_{i=1}^n$ which are independent random vectors distributed as X. Our aim is to estimate the deviation of operators of the form $N^{-1}\sum_{i=1}^{N} X_i \otimes X_i$ from the average operator $\mathbb{E}(X \otimes X)$, where $X \otimes X$ is the operator defined by $(X \otimes X)(v) = \langle X, v\rangle X$. Our investigation is motivated by two seemingly unrelated questions concerning the eigenvalues of some random matrices.

First, let X be a random point selected from a convex symmetric body in $\mathbb{R}^n$ which is in isotropic position. By this we mean the following: let $K \subset \mathbb{R}^n$ be a convex and symmetric set (i.e. if $x \in K$ then $-x \in K$) with a non-empty interior. We say that K is in isotropic position if for any $t \in \mathbb{R}^n$,

$$\frac{1}{\mathrm{vol}(K)} \int_K |\langle t, x\rangle|^2 \, \mathrm{d}x = \|t\|^2, \tag{1.1}$$

where the volume and the integral are with respect to the Lebesgue measure on $\mathbb{R}^n$, and $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ respectively denote the scalar product and the norm in the Euclidean space ℓ_2^n or in the infinite-dimensional space ℓ_2. In other words, if one considers the normalized volume measure on K and X is a random vector with that distribution, then a body is in isotropic position if for every $t \in \mathbb{R}^n$, $\mathbb{E}|\langle X, t\rangle|^2 = \|t\|^2$. It is easy to verify that for every convex, symmetric set K in $\mathbb{R}^n$ with a non-empty interior, there is some $T \in GL_n(\mathbb{R})$ such that TK is

1350–7265

isotropic. Note that we use a slightly different normalization than the standard definition of the isotropic position used in asymptotic geometry (for the more standard notion, see Milman and Pajor 1989), but for our purposes (1.1) is the correct normalization.

Consider the random operator $\Gamma : \mathbb{R}^n \to \mathbb{R}^N$ defined by

$$\Gamma = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{pmatrix},$$

where $(X_i)_{i=1}^N$ are independent random variables distributed according to the normalized volume measure on the body K. The ability to bound the largest and the smallest singular values of Γ has several applications (which will be explored in Section 3). For example, it is natural to ask whether the largest singular value of Γ is of the order $\sqrt{N}$ with high probability. Unfortunately, it seems that the standard method of estimating the largest eigenvalue fails here, unless one has much more information on the geometry of the body than the fact that it is in isotropic position.

For example, if the body has the property that for some constant c and every $t \in \mathbb{R}^n$ the random variable $Z_t = \langle X, t \rangle$ has a sub-Gaussian tail (i.e. $\Pr(\{|Z_t| \geqslant u\}) \leqslant 2\exp(-cu^2/\|t\|^2)$), then one can show using a standard ε-net argument that the largest singular value of Γ is indeed of the order $\sqrt{N}$. Unfortunately, most bodies do not exhibit this sub-Gaussian behaviour (for a characterization of such bodies, see Paouris 2005), and thus one must resort to a different approach to obtain an estimate on the largest singular value of Γ.

A difficulty arises because the matrix Γ has dependent entries, whereas in the standard set-up in the theory of random matrices, one investigates matrices with independent, identically distributed entries.

The method we use to address this problem is surprisingly simple. Note that if $N \geqslant n$, the first n eigenvalues of $\Gamma\Gamma^* = (\langle X_i, X_j \rangle)_{i,j=1}^N$ are the same as the eigenvalues of $\Gamma^*\Gamma = \sum_{i=1}^N X_i \otimes X_i$. We will show that under very mild assumptions on X, with high probability,

$$\left\| \frac{1}{N} \sum_{i=1}^N X_i \otimes X_i - \Lambda \right\|_{\ell_2^n \to \ell_2^n} \tag{1.2}$$

tends to 0 quickly as N tends to infinity, where $\Lambda = \mathbb{E}(X \otimes X)$, and we provide quantitative bounds on the rate of convergence and the 'high probability'. In particular, with high probability the eigenvalues of $N^{-1}\sum_{i=1}^N X_i \otimes X_i$ are close to the eigenvalues of Λ.

This general approximation question was motivated by an application in complexity theory, investigated by Kannan *et al.* (1997), regarding algorithms which approximate the volume of convex bodies. Previous results in the direction of estimating (1.2) were obtained by Bourgain and by Rudelson. Bourgain (1999) proved the following theorem:

Theorem 1.1. *For every $\varepsilon > 0$, there exists a constant $c(\varepsilon)$ for which the following holds. If K*

is a convex symmetric body in $\mathbb{R}^n$ *in isotropic position and* $N \geqslant c(\varepsilon) n \log^3 n$*, then with probability at least* $1-\varepsilon$*, for any* $t \in S^{n-1}$*,*

$$1-\varepsilon \leqslant \frac{1}{N}\sum_{i=1}^{N}\langle X_i, t\rangle^2 = \frac{1}{N}\|\Gamma t\|^2 \leqslant 1+\varepsilon.$$

Giannopoulos and Milman (2000) showed that Bourgain's method can actually give a better estimate of $N \geqslant c(\varepsilon) n \log^2 n$.

Equivalently, the previous inequalities say that $N^{-1/2}\Gamma\colon \ell_2^n \to \ell_2^N$ is a good embedding of ℓ_2^n.

Remark. When the random vector $X = (g_i)_{i=1}^n$ where $(g_i)_{i=1}^n$ are independent, standard Gaussian variables, it is known that for any $t \in S^{n-1}$,

$$1-2\sqrt{\frac{n}{N}} \leqslant \frac{1}{N}\sum_{i=1}^{N}\langle X_i, t\rangle^2 \leqslant 1+2\sqrt{\frac{n}{N}}$$

holds with high probability (see Davidson and Szarek 2001: Theorem II.13). This implies that in the Gaussian case, Theorem 1.1 is true for $N \geqslant 4n/\varepsilon^2$, and that this estimate is asymptotically optimal, up to a numerical constant.

Bourgain's result was improved by Rudelson (1999), who removed one power of the logarithm while proving a more general statement:

Theorem 1.2. *There exists an absolute constant C for which the following holds. Let Y be a random vector in* $\mathbb{R}^n$ *such that* $\mathbb{E}(Y \otimes Y) = Id$*. Then*

$$\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N} Y_i \otimes Y_i - Id\right\|_{\ell_2^n \to \ell_2^n} \leqslant C\sqrt{\frac{\log n}{N}}(\mathbb{E}\|Y\|^{\log N})^{1/\log N}.$$

Our main result, which is a deviation estimate for (1.2), implies the result of Rudelson, and its proof follows a similar path to his work.

The second application we present has a different flavour. Let $\Omega \subset \mathbb{R}^d$ and set ν to be a probability measure on Ω. Let t be a random variable on Ω distributed according to ν and set $X(t) = \sum_{i=1}^{\infty}\sqrt{\lambda_i}\phi_i(t)\phi_i$, where $(\phi_i)_{i=1}^{\infty}$ is an orthonormal basis in $L_2(\Omega, \nu)$ and $(\lambda_i)_{i=1}^{\infty} \in \ell_1$.

This choice of $X(t)$ originates in a question in nonparametric statistics which we now formulate. Let $L : \Omega \times \Omega \to \mathbb{R}$ be a bounded, positive-definite kernel. Under mild assumptions, by Mercer's theorem, there is an orthonormal basis of $L_2(\Omega, \nu)$, denoted by $(\phi_i)_{i=1}^{\infty}$, such that $\nu \otimes \nu$-almost surely, $L(t, s) = \sum_{i=1}^{\infty}\lambda_i\phi_i(t)\phi_i(s)$. Hence, $\langle X(t), X(s)\rangle = L(s, t)$ and the squares of the singular values of the random matrix Γ are the eigenvalues of the Gram matrix $(\langle X(t_i), X(t_j)\rangle)_{i,j=1}^N$, where $t_1, \ldots, t_N$ are independent random variables distributed according to ν. It is natural to ask whether the eigenvalues of this Gram matrix converge in some sense to the eigenvalues of the integral operator $T_L = \int L(x, y) f(y) \mathrm{d}\nu(y)$. This question was explored by Koltchinskii (1998) and

Koltchinskii and Giné (2000) and some partial results were obtained on the expected distance (with respect to the distance $d(x, y) = \inf_\sigma (\sum_{i=1}^{\infty}(x_i - y_{\sigma(i)})^2)^{1/2}$, with the infimum taken with respect to all permutations) between the set of empirical eigenvalues and the set of eigenvalues of the integral operator.

The significance of this question is that the eigenvalues of the integral operator play a key role in the analysis of kernel-based methods, often used in various statistical applications (for some theoretical results in this context, see Mendelson 2003 and references therein) but it is not clear how those should be estimated from the given data in the form of the Gram matrix. Our results enable us to do just that; indeed, if $X(t) = \sum_{i=1}^{\infty}\sqrt{\lambda_i}\phi_i(t)\phi_i$ then

$$\mathbb{E}(X \otimes X) = \sum_{i=1}^{\infty} \lambda_i \langle \phi_i, \cdot \rangle \phi_i = T_L.$$

Hence, a deviation inequality for $\|N^{-1}\sum_{i=1}^{N} X_i \otimes X_i - \Lambda\|_{\ell_2 \to \ell_2}$ enables one to estimate with high probability the eigenvalues of the integral operators using the eigenvalues of the Gram matrix.

We end this introduction with a notational convention. Throughout, all absolute constants are positive and will be denoted by C or c. Their values may change from line to line, or even within the same line. By $\|\cdot\|$ we denote either the ℓ_2 norm or the operator norm between ℓ_2 spaces. Other norms we use will be clearly specified.

2. The deviation inequality

Our starting point is the definition of the family of Orlicz norms. Recall that for a random variable Y and $\alpha \geqslant 1$, the ψ_α norm of Y is

$$\|Y\|_{\psi_\alpha} = \inf\left\{ C > 0;\ \mathbb{E}\exp\left(\frac{|Y|^\alpha}{C^\alpha}\right) \leqslant 2 \right\}.$$

A standard argument (de la Peña and Giné 1999; van der Vaart and Wellner 1996) shows that if Y has a bounded ψ_α norm then the tail of Y decays faster than $2\exp(-u^\alpha/\|Y\|_{\psi_\alpha}^\alpha)$. Moreover, a straightforward computation shows that for every $\alpha \geqslant 1$, if, for every integer $p \geqslant 1$, $(\mathbb{E}|Y^p|)^{1/p} \leqslant Kp^{1/\alpha}$, then $\|Y\|_{\psi_\alpha} \leqslant c_\alpha K$, where c_α is a constant which depends only on α.

Let us turn to the assumptions we need to make on the random vector X.

Assumption 2.1. *Let X be a random vector on $\mathbb{R}^n$ (or ℓ_2). We will assume the following:*

(i) *There is some $\rho > 0$ such that for every θ of norm* 1, $(\mathbb{E}|\langle X, \theta\rangle|^4)^{1/4} \leqslant \rho$.
(ii) *Set* $Z = \|X\|$. *Then* $\|Z\|_{\psi_\alpha} < \infty$ *for some* $\alpha \geqslant 1$.

Observe that Assumption 2.1 implies that the average operator Λ satisfies $\|\Lambda\| \leqslant \rho^2$. Indeed, denoting by S the sphere in either ℓ_2^n or ℓ_2,

$$\|\Lambda\| = \sup_{\theta_1,\theta_2 \in S} \langle \Lambda\theta_1, \theta_2 \rangle = \sup_{\theta_1,\theta_2 \in S} \mathbb{E}\langle X, \theta_1 \rangle \langle X, \theta_2 \rangle$$

$$\leqslant \sup_{\theta \in S} \mathbb{E}\langle X, \theta \rangle^2 \leqslant \rho^2.$$

The main result we shall establish is the following:

Theorem 2.1. *There exists an absolute constant c for which the following holds. Let X be a random vector in $\mathbb{R}^n$ (or ℓ_2) which satisfies Assumption 2.1 and set $Z = \|X\|$. For any integers n and N let $d = \min\{n, N\}$ if X is essentially supported in a finite-dimensional space and $d = N$ otherwise. If*

$$A_{d,N} = \|Z\|_{\psi_\alpha} \frac{\sqrt{\log d}(\log N)^{1/\alpha}}{\sqrt{N}} \quad \text{and} \quad B_{d,N} = \frac{\rho^2}{\sqrt{N}} + \|\Lambda\|^{1/2} A_{d,N}$$

then, for any $x > 0$,

$$\Pr\left(\left\| \sum_{i=1}^N (X_i \otimes X_i - \Lambda) \right\| \geqslant xN \right) \leqslant \exp\left[-\left(\frac{cx}{\max\{B_{d,N}, A_{d,N}^2\}} \right)^\beta \right],$$

where $\beta = (1 + 2/\alpha)^{-1}$ and $\Lambda = \mathbb{E}(X \otimes X)$.

As we show below, the probability we wish to estimate is the tail of the supremum of a centred empirical process. It is impossible to use standard concentration results for such processes (for example, Talagrand's inequality; see Ledoux 2001) because the indexing class of functions at hand is not bounded in L_∞.

The first step in the proof of Theorem 2.1 is a well-known symmetrization theorem (van der Vaart and Wellner 1996) which originated in the works of Kahane and Hoffman-Jørgensen. Recall that a Rademacher random variable is a random variable taking values ± 1 with probability $\frac{1}{2}$.

Theorem 2.2. *Let Z be a stochastic process indexed by a set F and let N be an integer. For every $i \leqslant N$, let $\mu_i : F \to \mathbb{R}$ be arbitrary functions and set $(Z_i)_{i \leqslant N}$ to be independent copies of Z. Under mild topological conditions on F and (μ_i) ensuring the measurability of the events below, for any $x > 0$,*

$$\beta_N(x) \Pr\left(\sup_{f \in F} \left| \sum_{i=1}^N Z_i(f) \right| > x \right) \leqslant 2\Pr\left(\sup_{f \in F} \left| \sum_{i=1}^N \varepsilon_i (Z_i(f) - \mu_i(f)) \right| > \frac{x}{2} \right),$$

where $(\varepsilon_i)_{i=1}^N$ are independent Rademacher random variables and

$$\beta_N(x) = \inf_{f \in F} \Pr\left(\left| \sum_{i=1}^N Z_i(f) \right| < \frac{x}{2} \right).$$

Observe that it is possible to express the operator norm of $\sum_{i=1}^N (X_i \otimes X_i - \Lambda)$ as the supremum of an empirical process. Indeed, let $\mathcal{U}$ be the set of tensors $v \otimes w$, where v and w are vectors in the unit Euclidean ball (or the unit ball in ℓ_2). Then

$$\|X \otimes X - \Lambda\| = \sup_{U \in \mathcal{U}} \langle X \otimes X - \Lambda, U \rangle,$$

where $\langle X \otimes X, v \otimes x \rangle = \langle X, v \rangle \langle X, w \rangle$.

Consider the process indexed by $\mathcal{U}$ defined by

$$Z(U) = \frac{1}{N} \sum_{i=1}^{N} \langle X_i \otimes X_i - \Lambda, U \rangle.$$

Clearly, for every U, $\mathbb{E} Z(U) = 0$ and

$$\sup_{U \in \mathcal{U}} Z(U) = \left\| \frac{1}{N} \sum_{i=1}^{N} (X_i \otimes X_i - \Lambda) \right\|.$$

Hence, to apply Theorem 2.2 one has to estimate, for any fixed $U \in \mathcal{U}$,

$$\Pr\left(\left| \sum_{i=1}^{N} \langle X_i \otimes X_i - \Lambda, U \rangle \right| > Nx \right).$$

It is easy to verify that for any $U \in \mathcal{U}$,

$$\operatorname{var}(\langle X \otimes X - \Lambda, U \rangle) \leqslant \sup_{\theta \in S} \mathbb{E} |\langle X, \theta \rangle|^4 \leqslant \rho^4.$$

In particular, $\operatorname{var}(Z(U)) \leqslant \rho^4 / N$, implying by Chebyshev's inequality that

$$\beta_N(2x) \geqslant 1 - \frac{\rho^4}{Nx^2}.$$

Corollary 2.3. *Let X be a random vector in $\mathbb{R}^n$ (or ℓ_2) which satisfies Assumption 2.1 and let $X_1, \ldots, X_N$ be independent copies of X. Then,*

$$\Pr\left(\left\| \sum_{i=1}^{N} (X_i \otimes X_i - \Lambda) \right\| > xN \right) \leqslant 4 \Pr\left(\left\| \sum_{i=1}^{N} \varepsilon_i X_i \otimes X_i \right\| > \frac{xN}{2} \right),$$

provided that $x \geqslant c\sqrt{\rho^4/N}$, for some absolute constant c.

The next step is an estimate on the norm of the symmetric random variable $\sum_{i=1}^{N} \varepsilon_i X_i \otimes X_i$.

We apply the following result of Rudelson (1999), which builds on an inequality due to Lust-Piquard and Pisier (1991).

Theorem 2.4. *There exists an absolute constant c such that for any integers n and N, any $x_1, \ldots, x_N \in \mathbb{R}^n$ (or ℓ_2) and any $p \geqslant 1$,*

$$\left(\mathbb{E}\left\|\sum_{i=1}^{N}\varepsilon_i x_i\otimes x_i\right\|^p\right)^{1/p}\leqslant c\max\{\sqrt{\log d},\ \sqrt{p}\}\left\|\sum_{i=1}^{N}x_i\otimes x_i\right\|^{1/2}\max_{1\leqslant i\leqslant N}\|x_i\|,$$

where $(\varepsilon_i)_{i=1}^N$ *are independent Rademacher random variables and* $d=\min\{n, N\}$.

Remark. The reason why one can select $d=\min\{n, N\}$ is that, for every realization of the Rademacher random variables, the norm of the operator $\sum_{i=1}^N\varepsilon_i x_i\otimes x_i$ is determined on the span of $\{x_1, \dots, x_N\}$ which is a Euclidean space of dimension at most d.

Note that this moment inequality immediately leads to a ψ_2 estimate on the random variable $\|\sum_{i=1}^N\varepsilon_i x_i\otimes x_i\|$.

Corollary 2.5. *There exists an absolute constant c such that for any integers n and N, any* $x_1, \dots, x_N\in\mathbb{R}^n$ *(or* ℓ_2*) and any* $t>0$,

$$\Pr\left(\left\{\left\|\sum_{i=1}^{N}\varepsilon_i x_i\otimes x_i\right\|\geqslant t\right\}\right)\leqslant 2\exp\left(-\frac{t^2}{\Delta^2}\right),$$

where $\Delta=c\sqrt{\log d}\|\sum_{i=1}^N x_i\otimes x_i\|^{1/2}\max_{1\leqslant i\leqslant N}\|x_i\|$ *and* $d=\min\{n, N\}$.

We are now ready to prove the main deviation inequality:

Proof of Theorem 2.1. First, by a result due to Pisier (1983; see also van der Vaart and Wellner 1996), if T is a random variable with a bounded ψ_α norm for $\alpha\geqslant 1$ and if $T_1, \dots, T_N$ are independent copies of T, then

$$\left\|\max_{1\leqslant i\leqslant N}T_i\right\|_{\psi_\alpha}\leqslant C\|T\|_{\psi_\alpha}\log^{1/\alpha}N$$

for an absolute constant C. Hence, for any integer p,

$$\left(\mathbb{E}\max_{1\leqslant i\leqslant N}|T_i|^p\right)^{1/p}\leqslant Cp^{1/\alpha}\|T\|_{\psi_\alpha}\log^{1/\alpha}N. \tag{2.1}$$

Consider the random variables

$$S=\left\|\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i X_i\otimes X_i\right\|\quad\text{and}\quad V=\left\|\frac{1}{N}\sum_{i=1}^{N}(X_i\otimes X_i-\Lambda)\right\|.$$

It follows from Corollaries 2.3 and 2.5 that, for any $t\geqslant c\sqrt{\rho^4/N}$,

$$\begin{aligned}\Pr(V\geqslant t)&\leqslant 4\Pr(S\geqslant t/2)=4\mathbb{E}_X\Pr_\varepsilon(S\geqslant t/2|X_1, \dots, X_N)\\&\leqslant 8\mathbb{E}_X\exp\left(-\frac{t^2N^2}{\Delta^2}\right),\end{aligned}$$

where $\Delta=c\sqrt{\log d}\|\sum_{i=1}^N X_i\otimes X_i\|^{1/2}\max_{1\leqslant i\leqslant N}\|X_i\|$ for some absolute constant c and

$d = \min\{n, N\}$. Setting c_0 to be the constant from Corollary 2.3, then by Fubini's theorem and dividing the region of integration into $t \leqslant c_0\sqrt{\rho^4/N}$ (in this range one has no control over $\Pr(V \geqslant t)$) and $t > c_0\sqrt{\rho^4/N}$, it is evident that

$$\begin{aligned}\mathbb{E}V^p &= \int_0^\infty pt^{p-1}\Pr(V \geqslant t)\mathrm{d}t \\ &\leqslant \int_0^{c_0\sqrt{\rho^4/N}} pt^{p-1}\,\mathrm{d}t + 8\mathbb{E}_X\int_0^\infty pt^{p-1}\exp\left(-\frac{t^2N^2}{\Delta^2}\right)\mathrm{d}t \\ &\leqslant (c_0\sqrt{\rho^4/N})^p + c^p p^{p/2}\,\mathbb{E}_X\left(\frac{\Delta}{N}\right)^p\end{aligned}$$

for some new absolute constant c.

The second term is bounded by

$$\begin{aligned}&c^p\left(\frac{p\log d}{N}\right)^{p/2}\mathbb{E}\left(\left\|\frac{1}{N}\sum_{i=1}^N X_i\otimes X_i\right\|^{p/2}\max_{1\leqslant i\leqslant N}\|X_i\|^p\right) \\ &\leqslant c^p\left(\frac{p\log d}{N}\right)^{p/2}\mathbb{E}\left(\left(\left\|\frac{1}{N}\sum_{i=1}^N X_i\otimes X_i-\Lambda\right\|+\|\Lambda\|\right)^{p/2}\max_{1\leqslant i\leqslant N}\|X_i\|^p\right) \\ &\leqslant c^p\left(\frac{p\log d}{N}\right)^{p/2}(\mathbb{E}(V+\|\Lambda\|)^p)^{1/2}\left(\mathbb{E}\max_{1\leqslant i\leqslant N}\|X_i\|^{2p}\right)^{1/2}\end{aligned}$$

for some other absolute constant c. Hence, setting $Z = \|X\|$ and applying Assumption 2.1 and (2.1), we arrive at

$$(\mathbb{E}V^p)^{1/p} \leqslant c\left(\frac{\rho^2}{\sqrt{N}} + p^{1/\alpha+1/2}\left(\frac{\log d}{N}\right)^{1/2}\log^{1/\alpha}N\|Z\|_{\psi_\alpha}\left((\mathbb{E}V^p)^{1/p}+\|\Lambda\|\right)^{1/2}\right),$$

for some absolute constant c. Set $A_{d,N} = (N^{-1}\log d)^{1/2}(\log^{1/\alpha}N)\|Z\|_{\psi_\alpha}$ and $\beta = (1+2/\alpha)^{-1}$. Thus,

$$(\mathbb{E}V^p)^{1/p} \leqslant c\left(\frac{\rho^2}{\sqrt{N}} + p^{1/2\beta}\|\Lambda\|^{1/2}A_{d,N} + p^{1/2\beta}A_{d,N}(\mathbb{E}V^p)^{1/2p}\right),$$

from which it is evident that

$$(\mathbb{E}V^p)^{1/p} \leqslant cp^{1/\beta}\max\left\{\frac{\rho^2}{\sqrt{N}} + \|\Lambda\|^{1/2}A_{d,N},\, A_{d,N}^2\right\},$$

and the assertion of the theorem follows. □

Let us present two corollaries which are relevant to the applications we have in mind. First, consider the case when ar X is a bounded random vector. Thus, for any α,

$\|Z\|_{\psi_\alpha} \leqslant \sup\|X\| \equiv \mathrm{R}$, and by taking $\alpha \to \infty$ one can select $\beta = 1$ and $A_{d,N} = R\sqrt{N^{-1}\log d}$, and obtain the following:

Corollary 2.6. *There exists an absolute constant c for which the following holds. Let X be a random vector in $\mathbb{R}^n$ (or ℓ_2) bounded by R and satisfying Assumption 2.1. Then, for any $x > 0$,*

$$\Pr\left(\left\{\left\|\sum_{i=1}^{N}(X_i \otimes X_i - \Lambda)\right\| \geqslant xN\right\}\right) \leqslant \exp\left(-\frac{cx}{R^2}\min\left\{\frac{\sqrt{N}}{\sqrt{\log d}}, \frac{N}{\log d}\right\}\right).$$

The second case is when X is a vector on $\mathbb{R}^n$ and $\|Z\|_{\psi_2} \leqslant c\sqrt{n}$, which corresponds to the geometric application we have in mind, where X is a random vector associated with a convex body in isotropic position.

Corollary 2.7. *There exists an absolute constants c for which the following holds. Let X be a random vector in $\mathbb{R}^n$ which satisfies Assumption 2.1 with $\|Z\|_{\psi_2} \leqslant c_1\sqrt{n}$. Then, for any $x > 0$,*

$$\Pr\left(\left\{\left\|\sum_{i=1}^{N}(X_i \otimes X_i - \Lambda)\right\| \geqslant xN\right\}\right)$$
$$\leqslant \exp\left(-c\left(x/\max\left\{\frac{\rho^2}{\sqrt{N}} + \rho c_1\sqrt{\frac{(n\log n)\log N}{N}}, c_1^2\frac{(n\log n)\log N}{N}\right\}\right)^{1/2}\right).$$

3. Applications

The first application we present is when the random variable X corresponds to the volume measure of some convex symmetric body in isotropic position, which fits our assumptions perfectly. Indeed, as shown by Alesker (1995), there exists an absolute constant C such that if K is in isotropic position and if $Z = \|X\|$ then $\|Z\|_{\psi_2} \leqslant C\sqrt{n}$. Moreover, by the Brunn–Minkowski inequality, if K is in isotropic position then its volume measure is log-concave and $\mathbb{E}|\langle X, t\rangle|^2 = \|t\|^2$ for any $t \in \mathbb{R}^n$. Hence, if $\theta \in S^{n-1}$ then $\Pr(\{|\langle X, \theta\rangle| \geqslant 2\}) \leqslant \frac{1}{4}$, and by Borell's inequality, $\|\langle X, \theta\rangle\|_{\psi_1} \leqslant C$ for some new absolute constant (see Ledoux 2001; Milman and Schechtman 1986). In particular, Assumption 2.1 is verified for $\alpha = 2$, $\|Z\|_{\psi_2} \leqslant C\sqrt{n}$ and $\rho = C$, which is the situation in Corollary 2.7. Moreover, by the definition of the isotropic position, $\mathbb{E}(X \otimes X) = Id$.

Let us note a few simple outcomes of these observations; similar results can be obtained equally easily.

Corollary 3.1. *There are absolute constants c_1, c_2, c_3, c_4 for which the following holds. Let*

$K \subset \mathbb{R}^n$ be a symmetric convex body in isotropic position, let $X_1, \ldots, X_N$ be independent points sampled according the normalized volume measure on K, and set

$$\Gamma = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{pmatrix}$$

with non-zero singular values $\lambda_1, \ldots, \lambda_n$.

(i) *If $N \geqslant c_1 n \log^2 n$, then for every $x \geqslant 0$,*

$$\begin{aligned} &\Pr(\{\forall i,\ (1-x)^{1/2}\sqrt{N} \leqslant \lambda_i \leqslant (1+x)^{1/2}\sqrt{N}\}) \\ &\geqslant 1 - \exp\left(-c_2 x^{1/2}\left(\frac{N}{(\log N)(n\log n)}\right)^{1/4}\right). \end{aligned}$$

(ii) *If $N \geqslant c_3 n$, then with probability at least $\frac{1}{2}$, $\lambda_1 \leqslant c_4\sqrt{N\log n}$.*

In particular, this estimate shows that if N is polynomial in n one can get the correct estimate on the largest singular value of Γ, and with relatively high probability. Also, as long as $N \geqslant cn\log^2 n$, the bound on all singular values is non-trivial. With as little as $N \sim n^5 \log^2 n$ one obtains that the same holds with probability $\exp(-cn)$, which complements a result from Litvak *et al.* (2005) but is, most likely, a suboptimal estimate.

The fact that one can bound the smallest singular value has a geometric interpretation, as it implies that the symmetric convex hull of $X_1, \ldots, X_N$ contains a 'large' Euclidean ball. Indeed, let A be the symmetric convex hull of $\{X_1, \ldots, X_N\}$. By duality, it is easy to verify that $rB_2^n \subset A$ if and only if $r\|x\|_2 \leqslant \|\Gamma x\|_\infty$ for every $x \in \mathbb{R}^n$. Hence it suffices to show that $r\sqrt{N}\|x\|_2 \leqslant \|\Gamma x\|_2$, which is a condition on the smallest singular value of Γ.

Corollary 3.2. *Let K be as in Corollary 3.1 and let A be the symmetric convex hull of $X_1, \ldots, X_N$. Then, for every $0 < \varepsilon < \frac{1}{2}$ and integers $N \geqslant n$,*

$$\begin{aligned} &\Pr(\{(1-\varepsilon)B_2^n \subset A\}) \\ &\geqslant 1 - \exp\left(-C\varepsilon\left(\min\left\{\sqrt{\frac{N}{(\log N)(n\log n)}}, \frac{N}{(\log N)(n\log n)}\right\}\right)^{1/2}\right). \end{aligned}$$

In particular, if $N \geqslant c(\varepsilon)n\log^2 n$ then $(1-\varepsilon)B_2^n \subset A$ with probability larger than $\frac{1}{2}$.

Note that the 'in particular' part also follows from Rudelson's result (Theorem 1.2), but it does not imply the better concentration if one takes $N \gg cn\log^2 n$.

This estimate is almost optimal in the following sense: by the log-concavity of the volume measure on K, combined with Borell's inequality (see Ledoux 2001; Milman and Schechtman 1986),

$$\Pr(\{X \notin K \cap cr\sqrt{n}B_2^n\}) \leqslant c^{(1+r)/2}$$

for some $c < 1$. Hence, with high probability, $X_1, \ldots, X_N \in c \log N \cdot \sqrt{n}B_2^n$, and by the Carl–Pajor inequality (Carl and Pajor 1988; see also Bárány and Füredy 1988; Gluskin 1988) for $N \sim n \log^2 n$,

$$\mathrm{vol}^{1/n}(\mathrm{absconv}\ (X_1, \ldots, X_N)) \leqslant c \log N \cdot \sqrt{\frac{\log(N/n)}{n}} \leqslant c \frac{\log n \cdot (\log \log n)^{1/2}}{\sqrt{n}},$$

while with probability at least $\frac{1}{2}$,

$$\mathrm{vol}^{1/n}(\mathrm{absconv}\ (X_1, \ldots, X_N)) \geqslant \frac{c}{\sqrt{n}},$$

because the symmetric convex hull contains a ball of radius $\frac{1}{2}$.

The second application we present deals with the eigenvalues of integral operators. Let L be a positive-definite kernel on some probability space (Ω, μ). Assume that L is continuous and that Ω is compact. Thus, by Mercer's theorem, $L(x, y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(y)$, where $(\lambda_i)_{i=1}^{\infty}$ are the eigenvalues of T_L, the integral operator associated with L and μ, and $(\phi_i)_{i=1}^{\infty}$ is a complete orthonormal basis in $L_2(\mu)$. Also, T_L is a trace-class operator, since $\sum_{i=1}^{\infty} \lambda_i = \int L(x, x) \mathrm{d}\mu(x)$.

Let $X(t) = \sum_{i=1}^{\infty} \sqrt{\lambda_i} \phi_i(t) \phi_i \in L_2$. We would like to apply Theorem 2.1 to this vector for two reasons. First, observe that $\mathbb{E}(X \otimes X)$ is the integral operator T_L and for every $t_1, \ldots, t_N$, the squares of the singular values of the matrix Γ are the eigenvalues of the Gram matrix $\left(L(t_i, t_j)\right)_{i,j=1}^{N}$. Hence, a successful application of Theorem 2.1 would yield a deviation inequality between the eigenvalues of the Gram matrix and those of the integral operator.

The second reason uses the whole power of the approximation result. In some applications in nonparametric statistics, one sometimes has the fixed mapping $t \rightarrow X(t)$ without having additional information on the Mercer representation of the integral operator (e.g. if the eigenfunctions are not known). Our result enables one to approximate the integral operator using a finite-dimensional approximation in such cases (of course, if one has the Mercer representation of L, finding such a finite-dimensional approximation is trivial).

Observe that $\|X(t)\|^2 = \langle X(t), X(t) \rangle = L(t, t)$. Hence, if we set $R = \|L(t, t)\|_{\infty}^{1/2}$ then $\|X(t)\| \leqslant R$. Also, recall that for compact, self-adjoint operators $A, B : \ell_2 \rightarrow \ell_2$, $\sup_i |\lambda_i(A) - \lambda_i(B)| \leqslant \|A - B\|$, where $(\lambda_i(A))$ denotes the sequence of the singular values of the operator A arranged in a non-increasing order (Gohberg and Goldberg 1980). Applying this fact to the operators $A = N^{-1} \sum_{i=1}^{N} X_i \otimes X_i$ and $B = \mathbb{E}(X \otimes X)$, then by Corollary 2.6 we obtain the following theorem:

Theorem 3.3. *There exists an absolute constant c for which the following holds. Let L be as above and let* $\hat{\lambda}_1 \geqslant \ldots \geqslant \hat{\lambda}_N$ *be the eigenvalues of the Gram matrix* $(L(t_i, t_j))_{i,j=1}^{N}$. *Then, for every* $x > 0$,

$$\Pr\left(\sup_i |\hat{\lambda}_i - \lambda_i| \geqslant x\right) \leqslant 2\exp\left(-\frac{cx}{\|L(t,\, t)\|_\infty}\sqrt{\frac{N}{\log N}}\right),$$

where $(\lambda_i)_{i=1}^{\infty}$ *are the eigenvalues of the integral operator* T_L *and, for* $i > N$, $\hat{\lambda}_i = 0$.

Acknowledgement

The authors were partially supported by an Australian Research Council grant.

References

Alesker, S. (1995) ψ_2 estimates for the Euclidean norm on a convex body in isotropic position. *Oper. Theory Adv. Appl.*, **77**, 1–4.

Bárány, I. and Füredy, Z. (1988) Approximation of the sphere by polytopes having few vertices. *Proc. Amer. Math. Soc.*, **102**(3), 651–659.

Bourgain, J. (1999) Random points in isotropic convex bodies. In K.M. Ball and V. Milman (eds) *Convex Geometric Analysis*, Math. Sci. Res. Inst. Publ. 34, pp. 53–58. Cambridge: Cambridge University Press.

Carl, B. and Pajor, A. (1988) Gelfand numbers of operators with values in a Hilbert space. *Invent. Math.*, **94**, 479–504.

Davidson, K. and Szarek, S. (2001) Local operator theory, random matrices and Banach spaces. In W.B. Johnson and J. Lindenstrauss (eds), *Handbook of the Geometry of Banach Spaces*, Vol. I, pp. 317–366. Amsterdam: Elsevier.

de la Peña, V. and Giné, E. (1999) *Decoupling*. New York: Springer-Verlag.

Giannopoulos, A.A. and Milman, V.D. (2000) Concentration property on probability spaces. *Adv. Math.*, **156**(1), 77–106.

Gluskin, E.D. (1988) Extremal properties of orthogonal parallelepipeds and their applications to the geometry of Banach spaces. *Mat. Sb.* (N.S.) **136**(178), 85–96; translation in *Math. USSR Sb.* 64(1), 85–96 (1989).

Gohberg, I. and Goldberg, G. (1980) *Basic Operator Theory*. Basel: Birkhäuser.

Kannan, R., Lovász, L. and Simonovits, M. (1997) Random walks and $O^*(n^5)$ volume algorithm for convex bodies. *Random Structures Algorithms*, **2**(1), 1–50.

Koltchinskii, V. (1998) Asymptotics of spectral projections of some random matrices approximating integral operators. In E. Eberlein, M. Hahn and M. Talagrand (eds), *High Dimensional Probability*, Progr. Probab. 43, pp. 191–227. Basel: Birkhäuser.

Koltchinskii, V. and Giné, E. (2000) Random matrix approximation of spectra of integral operators. *Bernoulli*, **6**, 113–167.

Ledoux, M. (2001) *The Concentration of Measure Phenomenon*, Math. Surveys Monogr. 89. Providence, RI: American Mathematical Society.

Litvak, A., Pajor, A., Rudelson, M. and Tomczak-Jaegermann, N. (2005) Smallest singular value of random matrices and geometry of random polytopes. *Adv. Math.*, **195**, 491–523.

Lust-Piquard, F. and Pisier, G. (1991) Non-commutative Khinchine and Paley inequalities. *Ark. Mat.*, **29**, 241–260.

Mendelson, S. (2003) On the performance of kernel classes. *J. Mach. Learn. Res.*, **4**, 759–771.

Milman, V.D. and Pajor, A. (1989) Isotropic position and inertia ellipsoids and zonoids of the unit ball of a normed n-dimensional space. In J. Lindenstrauss and V.D. Milman (eds) *Geometric Aspects of Functional Analysis*, Lecture Notes in Math. 1376, pp. 64–104. Berlin: Springer-Verlag.

Milman, V.D. and Schechtman, G. (1986) *Asymptotic Theory of Finite Dimensional Normed Spaces*, Lecture Notes in Math. 1200. Berlin: Springer-Verlag.

Paouris, G. (2005) On the ψ_2 behaviour of linear functionals on isotropic convex bodies. *Studia Math.*, **168**, 285–299.

Pisier, G. (1983) Some applications of the metric entropy condition to harmonic analysis. In R.C. Blei and S.J. Sidney (eds), *Banach Spaces, Harmonic Analysis, and Probability Theory*, Lecture Notes in Math. 995, pp. 123–154. Berlin: Springer-Verlag.

Rudelson, M. (1999) Random vectors in the isotropic position. *J. Funct. Anal.*, **164**, 60–72.

van der Vaart, A.W. and Wellner, J.A. (1996) *Weak Convergence and Empirical Processes*. New York: Springer-Verlag.

Received October 2004 and revised February 2006