Ellipsoid Approximation Using Random Vectors

S. Mendelson[1] and A. Pajor[2]

[1] Centre for Mathematics and its Applications, The Australian National University, Canberra, ACT 0200, Australia
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
pajor@math.univ-mlv.fr

Abstract. We analyze the behavior of a random matrix with independent rows, each distributed according to the same probability measure on $\mathbb{R}^n$ or on ℓ_2. We investigate the spectrum of such a matrix and the way the ellipsoid generated by it approximates the covariance structure of the underlying measure. As an application, we provide estimates on the deviation of the spectrum of Gram matrices from the spectrum of the integral operator.

1 Introduction

Our objective is to explore the behavior of random vectors in $\mathbb{R}^n$ (resp. ℓ_2), particularly in the context of kernel methods and kernel Principal component analysis. To be more exact, let us formulate the two questions that motivated this study (though are not necessarily the main focus here).

Question 1. *Let (Ω, μ) be a probability space and let $K : \Omega \times \Omega \to \mathbb{R}$ be a positive definite kernel. Set $T_K : L_2(\mu) \to L_2(\mu)$ to be the integral operator associated with K and μ, given by*

$$(T_K f)(t) = \int K(s,t) f(s) d\mu(s).$$

Let $t_1, ..., t_N$ be independent random variables distributed according to μ, and let $\hat{T} = \left(\frac{1}{N} K(t_i, t_j)\right)_{i,j=1}^N$ be the corresponding Gram matrix. Does the spectrum of $\hat{T}$ converge (in an appropriate sense) to the spectrum of T_K?

Question 1 was studied by Koltchinskii and Giné [8] for a very wide range of kernels. They showed, among other things, that if K is a finite dimensional kernel, then the spectrum of $\hat{T}$ converges to that of T in an appropriate sense as N tends to infinity, and obtained estimates on the rate of convergence, which we improve here. Let us mention that most of the effort in [8] was devoted to the study of kernels which are not trace class (that is, $\mathbb{E}K(t,t) = \infty$), for which additional arguments are required, and our results do not cover that situation.

P. Auer and R. Meir (Eds.): COLT 2005, LNAI 3559, pp. 429–443, 2005.

The second question is connected to kernel PCA [11, 4]. Let X be a random vector in ℓ_2 (that is, a function from the probability space (Ω, μ) to ℓ_2), and let $X_1, ..., X_N$ be independent copies of X. Set $\{e_1, ..., e_N\}$ to be the standard unit basis in the N-dimensional Euclidean space ℓ_2^N and put $\mathcal{K}$ to be the image of the N-dimensional Euclidean ball, B_2^N, by the random operator $\Gamma : \ell_2^N \to \ell_2$ defined by $\Gamma e_i = X_i$.

For $d \leq N$, let $a_d = \inf \{\sup_{x \in \mathcal{K}} d(x, E) : \ E \subset \ell_2, \ \dim(E) = d\}$, that is, a_d is the best degree of approximation by which a d-dimensional subspace approximates the random ellipsoid $\mathcal{K}$.

Question 2. *Let E_d be the best approximating d-dimensional subspace as above. What estimates can one provide on the random variable $d(E_d, X_{N+1})$?*

In other words, the question is how close X_{N+1} is to the d-dimensional subspace that best approximates ΓB_2^N. Although we do not tackle this problem directly here, we present a method of attack which should be explored further, as explained below.

It turns out that both these questions are connected to the structure of random ellipsoids. Indeed, if $X(t)$ is a random vector in ℓ_2, it can be used to define a new Euclidean structure on its span, given by

$$\|v\|^2 = \mathbb{E} \left| \langle X, v \rangle \right|^2 . \tag{1.1}$$

Since this norm is given by the inner product $[u, v] = \mathbb{E} \langle X, u \rangle \langle X, v \rangle$, its unit ball is an ellipsoid (usually called the *Binet ellipsoid*) and is denoted by $\mathcal{E}_B$.

As an example, consider the integral operator T_K. Under mild assumptions on K and Ω, by Mercer's Theorem, there is an orthonormal basis of L_2, denoted by $(\phi_i)_{i=1}^{\infty}$, such that $K(s,t) = \sum_{i=1}^{\infty} \lambda_i \phi_i(s) \phi_i(t)$ almost surely, where $(\lambda_i)_{i=1}^{\infty}$ are the eigenvalues of the integral operator T_K arranged in a non-increasing order (in fact, for our needs it is enough that the convergence is in the L_2 sense rather than almost surely, for which it suffices that K is a positive definite, square integrable kernel, and we will make these assumptions on K throughout this note). Let $X(t) = \sum_{i=1}^{\infty} \sqrt{\lambda_i} \phi_i(t) \phi_i \in \ell_2$ (here we identify $L_2(\mu)$ with ℓ_2 and (ϕ_i) with the standard basis in ℓ_2), and consider the ellipsoid

$$\mathcal{E} = \left\{ v \in \ell_2 \ : \ \sum_{i=1}^{\infty} \frac{\langle v, \phi_i \rangle^2}{\lambda_i} \leq 1 \right\}.$$

Hence, $\mathcal{E}$ is an ellipsoid with principal directions ϕ_i and the "principal lengths" are $\sqrt{\lambda_i}$. We define the polar body of $\mathcal{E}$ by

$$\mathcal{E}^\circ = \{ y \in \ell_2 \ : \ \forall x \in \mathcal{E} \ \ | \langle x, y \rangle | \leq 1 \} .$$

Let us mention that the polar of a unit ball of some finite dimensional normed space X is the unit ball of the dual space X^*. Hence, in this case $\mathcal{E}^\circ$ is simply the unit ball of dual norm to the one defined by $\mathcal{E}$. Indeed, it is easy to verify that $\mathcal{E}^\circ$ is an ellipsoid, and with respect to the norm $\| \ \|_{\mathcal{E}^\circ}$, for which $\mathcal{E}^\circ$ is its unit ball,

$$\|v\|_{\mathcal{E}^\circ}^2 = \sum_{i=1}^{\infty} \lambda_i \langle v, \phi_i \rangle^2 = \mathbb{E}|\langle X(t), v \rangle|^2.$$

Thus, the Binet ellipsoid associated with $X(t)$ is the polar body of the ellipsoid $\mathcal{E}$.

In general, we define the ellipsoid $\mathcal{E}$ as the polar of $\mathcal{E}_B$. Both these ellipsoids are generated according to the covariance structure endowed by the random vector X.

Let $X_1, ..., X_N$ be independent copies of X, set $\Gamma_N : \ell_2^N \to \ell_2$ to be the random operator defined by $\Gamma_N e_i = \frac{1}{\sqrt{N}} X_i$ and denote $\hat{\mathcal{E}} = \Gamma_N B_2^N$. There are three natural questions that one can ask regarding various approximations of $\mathcal{E}$ using $\hat{\mathcal{E}}$.

1. How close are the lengths of the principal directions of $\hat{\mathcal{E}}$ to those of $\mathcal{E}$?
2. Is $\hat{\mathcal{E}}$ close to being a section of $\mathcal{E}$? (in other words, if $E = \text{span}\{X_1, ..., X_N\}$, is $\hat{\mathcal{E}}$ close to $\mathcal{E} \cap E$?
3. How close is $\hat{\mathcal{E}}$ to $\mathcal{E} \cap W_N$, where W_N is the subspace of ℓ_2 spanned by the N largest principal directions of $\mathcal{E}$?

Observe that understanding these questions would lead to answers to Question 1 and Question 2. Indeed, if $X(t) = \sum_{i=1}^{\infty} \sqrt{\lambda_i} \phi_i(t) \phi_i$ is generated by the kernel K, then the lengths of the principal directions of $\hat{\mathcal{E}}$ are the square roots of the eigenvalues of the matrix $\Gamma_N^* \Gamma_N$, which is the Gram matrix $\left(\frac{1}{N} K(t_i, t_j)\right)_{i,j=1}^N$. On the other hand, the principal lengths of $\mathcal{E}$ are $(\sqrt{\lambda_i})_{i=1}^{\infty}$. And thus, (1) for this specific choice of the random vector $X(t)$ is simply Question 1.

Next, suppose that the answer to (3) is affirmative, and the random ellipsoid $\hat{\mathcal{E}}$ approximates the section of $\mathcal{E}$ generated by the first N principal directions. Thus, the best d-dimensional approximating subspace is close to the space spanned by the first d principal directions of $\mathcal{E}$, implying that $d(E_d, X_{N+1}) \approx d(W_d, X_{N+1})$ which can be easily estimated.

It turns out that the degree of difficulty of (1)-(3) is increasing. Roughly speaking, (1) deals with the fact that $\hat{\mathcal{E}}$ is an ellipsoid which is close to a "rotation" of $\mathcal{E} \cap W_N$, as the principal lengths of $\hat{\mathcal{E}}$ are close to the N largest of $\mathcal{E}$. On the other hand, (2) identifies $\hat{\mathcal{E}}$ as being close to a section of $\mathcal{E}$, and depends on approximating both the principal lengths *and* the principal directions. Intuitively, (3) follows from a combination of (1) and (2) (under some mild assumptions on $\mathcal{E}$); if $\hat{\mathcal{E}}$ is almost a section of $\mathcal{E}$ and has the same principal lengths as the first N largest of $\mathcal{E}$, it must be close to $\mathcal{E} \cap W_N$.

Here, we will only investigate (2) and (3) when X is a vector in $\mathbb{R}^n$ and $N > n$. In this case we will show that $\hat{\mathcal{E}}$ is a good approximation of $\mathcal{E}$ rather than of a section of $\mathcal{E}$, and (2) and (3) coincide.

The main stumbling block in the study of the singular values of the random operator Γ_N which maps e_i to $X_i/\sqrt{N}$ (or, for that matter, the singular values of the Gram matrix $\left(\frac{1}{N} \langle X_i, X_j \rangle\right)_{i,j=1}^N$) is that the random matrix defined by this operator has dependent entries. One can bypass this problem by considering the operator $\Gamma\Gamma^* = \sum_{i=1}^N X_i \otimes X_i$ (where $X_i \otimes X_i$ is the projection onto the vector X_i, that is, for any $v \in \ell_2$, $(X_i \otimes X_i)(v) = \langle X_i, v \rangle X_i$). Observe that the N

largest eigenvalues of $\Gamma\Gamma^*$ are the same as the squares of the singular values of Γ and the advantage is that $\sum_{i=1}^N X_i \otimes X_i$ is a sum of independent, identically distributed, operator-valued variables. One can define the average operator (also known as the covariance operator) $\Lambda = \mathbb{E}(X \otimes X)$ as the operator which satisfies for any $u, v \in \ell_2$, $\langle \Lambda u, v\rangle = \mathbb{E}\langle (X \otimes X)u, v\rangle = \mathbb{E}\langle X, u\rangle \langle X, v\rangle$, and it is standard to verify that such an operator exists under mild integrability assumptions on X.

We will investigate the way the random operator $\frac{1}{N}\sum_{i=1}^N X_i \otimes X_i$ deviates from the average operator Λ with respect to various operator norms. Recall that for any normed space $(E, \|\ \|_E)$, if Y is a E-valued random variable, the process $\left\| \frac{1}{N}\sum_{i=1}^N (Y_i - \mathbb{E}Y)\right\|_E$ is the supremum of an empirical process which is indexed by the unit ball of the dual space of E. Thus, one can apply standard tools from empirical processes theory, such as symmetrization inequalities and concentration results. Let us point out that unlike most situations studied in Learning Theory, the random vector Y we deal with here need not be bounded; thus the class of functions defined by the dual unit ball is not uniformly bounded and Talagrand's concentration inequality for empirical processes indexed by bounded classes no longer applies.

Two types of assumptions are often used to compensate for the absence of an L_∞ bound on the class of functions. The first deals with the rate of decay of the linear forms $x^*(Y)$, and the other is on the rate of decay of the norm $\|Y\|_E$. To formulate these assumptions, let us recall the notion of Orlicz norms.

Definition 1. *For a random variable V and $\alpha \geq 1$, the ψ_α norm of V is*

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

A standard argument [14] shows that if V has a bounded ψ_α norm then its tail decays faster than $2\exp(-u^\alpha/\|V\|_{\psi_\alpha}^\alpha)$. In particular, a ψ_2 random variable has a subgaussian tail and a ψ_1 variable has a sub-exponential tail. If one assumes that the linear forms decay quickly, that is, are bounded with respect to an appropriate ψ_α norm, then using the *Generic Chaining method* [13], it is possible to upper bound the expectation of the supremum of an empirical process indexed by the dual unit ball B_{E^*}, using the metric structure of the space (B_{E^*}, ψ_α). We will not explore this direction here, but rather, formulate without a proof a relatively standard result which follows from this method.

Theorem 1. *For every $K > 0$ and $0 < \delta < 1$, there exist a constant $c(K, \delta)$ for which the following holds. Let X be a random vector in ℓ_2^n and let $\mathcal{E}_B$ be its Binet ellipsoid, which is assumed to have a full rank. If, for every $v \in \ell_2^n$, $\|\langle X, v\rangle\|_{\psi_2} \leq K(\mathbb{E}|\langle X, v\rangle|^2)^{1/2} = K\|v\|_{\mathcal{E}_B}$, then for any $0 < \varepsilon < 1$ and $N \geq c(K, \delta)n/\varepsilon^2$, with probability at least $1 - \delta$, every $v \in \mathbb{R}^n$ satisfies,*

$$(1-\varepsilon)\|v\|_{\mathcal{E}_B} \leq \left(\frac{1}{N}\sum_{i=1}^N \langle X_i, v\rangle^2\right)^{1/2} \leq (1+\varepsilon)\|v\|_{\mathcal{E}_B}. \tag{1.2}$$

Theorem 1 gives an equivalence between the ellipsoid $\hat{\mathcal{E}}^\circ$, which is the polar of $\Gamma_N B_2^N$, and the Binet ellipsoid. Unfortunately, such a result has several intrinsic limitations. First of all, the degree of approximation it provides is possibly too strong for our goals, in the following sense. Let $\lambda_1^{1/2}, ..., \lambda_n^{1/2}$ be the n (nonzero) singular values of Γ_N. Then, for any $v \in \mathbb{R}^n$, $\|v\|_{\mathcal{E}_B}^2 = \sum_{i=1}^n \lambda_i v_i^2$, and if many of the λ_is are very small, one can have vectors on the ℓ_2^n unit sphere, but with a small ellipsoid norm. Since Theorem 1 states that $\mathcal{E}_B \subset (1+\varepsilon)\hat{\mathcal{E}}^\circ$ and $\hat{\mathcal{E}}^\circ \subset (1+\varepsilon)\mathcal{E}_B$, its assertion is more restrictive than, say,

$$\mathcal{E}_B \subset \hat{\mathcal{E}}^\circ + \varepsilon B_2^n \quad \text{and} \quad \hat{\mathcal{E}}^\circ \subset \mathcal{E}_B + \varepsilon B_2^n, \tag{1.3}$$

where $A + B = \{a + b \ : \ a \in A, \ b \in B\}$. Equation (1.3) implies that each point in $\mathcal{E}_B$ can be written as a sum of a point in the random ellipsoid and a point with a small Euclidean norm and vice-versa, which would suffice in many applications.

The price one pays for the strong degree of approximation in Theorem 1 is that the bound holds only when the number of sample points N is of the order of the dimension n. And, there is no advantage if the singular values of Γ_N are small. This perhaps helps to explain the remark we made - that to see how well the random ellipsoid approximates the deterministic one is intrinsically more difficult if one selects this strong sense of approximation, because the fact that one has many "small" principal directions does not play to ones advantage.

The second problem with this approach is that the ψ_2 assumption on the linear forms $\langle X, v\rangle$ is very difficult to check, and is often not even true. In certain problems in convex geometry one can verify such an assumption, but in general, it is too much to hope for. Moreover, even in geometric scenarios, a more realistic assumption is a ψ_1 condition rather than a ψ_2 condition, which makes the analysis of the problem much more difficult, and Theorem 1 is no longer true as stated (see [2, 12, 3] for more details).

The approach we take here is to assume that probability that $\|X\|$ is large decays quickly (though $\|X\|$ need not be bounded) rather than the linear forms. In the context of integral operators, the motivation for this type of assumption is clear, since $\|X(t)\|_2^2 = K(t,t)$. Thus, one only has to consider the decay properties of the diagonal of the kernel. To that end, in most of the results we present, we require the following assumption:

Assumption 1. *Let X be a random vector in ℓ_2^n (resp. ℓ_2). Assume that*

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

In other words, the assumptions we make are on the fourth moment of linear forms $\langle X, \theta\rangle$, and (which is the more important part), on the decay properties of $\|X\|$. The first assumption follows if the second one is verified (and with essentially the same constant), using a Cauchy-Schwarz inequality and the fact that the L_p norm is upper bounded by the ψ_α norm, although in some cases one can obtain a better estimate on ρ.

1.1 Some Preliminaries

To derive tail estimates for $\left\| \frac{1}{N}\sum_{i=1}^N X_i \otimes X_i - \mathbb{E}(X \otimes X) \right\|_{2\to 2}$ (where $\| \ \|_{2\to 2}$ is the operator norm from ℓ_2 to ℓ_2), we shall use a well known symmetrization theorem [14] that originated in the works of Kahane and Hoffman-Jørgensen. Recall that a Rademacher random variable is a random variable taking values ± 1 with probability $1/2$.

Theorem 2. *Let Z be a stochastic process indexed by a set F and let N be an integer. For every $i \leq N$, let $\mu_i : F \to \mathbb{R}$ be arbitrary functions and set $(Z_i)_{i\leq N}$ to be independent copies of Z. Under mild topological conditions on F and (μ_i) ensuring the measurability of the events below, for any $x > 0$,*

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

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 in the case of the ℓ_2 operator norm, the supremum of an empirical process is taken with respect to $\mathcal{U}$ - the set of tensors $v \otimes w$, where v and w are vectors in the unit Euclidean ball, in which case, $\|X \otimes X - \Lambda\|_{2\to 2} = \sup_{U\in\mathcal{U}} \langle X \otimes X - \Lambda, U\rangle$. The next corollary follows from a standard estimate on $\beta_N(x)$, and its proof is omitted.

Corollary 1. *Let X be a random vector which satisfies Assumption 1 and let $X_1, ..., X_N$ be independent copies of X. Then,*

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

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

Thanks to the symmetrization argument and to the fact that for every empirical process

$$\mathbb{E} \sup_{f\in F} \left| \sum_{i=1}^N (f(X_i) - \mathbb{E}f) \right| \leq 2\mathbb{E}_{X\times\varepsilon} \sup_{f\in F} \left| \sum_{i=1}^N \varepsilon_i f(X_i) \right|,$$

it is enough to analyze the way operators of the form $\sum_{i=1}^N \varepsilon_i x_i \otimes x_i$ behave for a fixed set $x_1, ..., x_N$ in $\mathbb{R}^n$, or more generally, in ℓ_2.

Remark 1. *Observe that even if $x_i \in \ell_2$, in order to compute the operator norm of $\sum_{i=1}^N \varepsilon_i x_i \otimes x_i$, it suffices to restrict the operator to the span of $x_1, ..., x_N$, and thus we can assume that $x_i \in \ell_2^d$ for $d = \min\{N, n\}$.*

We will use several operator norms in what follows - all of which are connected to the singular values of an operator between two Hilbert spaces. The next definition is presented only in the finite dimensional case, but it has an obvious infinite dimensional analog.

Definition 2. *For $1 \leq p < \infty$, let C_p^d be the space of operators on $\mathbb{R}^d$, endowed with the norm $\|T\|_{C_p^n} = \left(\sum_{i=1}^n s_j^p(T)\right)^{1/p}$, where $s_j(T)$ is the j-th singular value of T. The space C_p^d is called the p-th Schatten class of $\mathbb{R}^d$.*

Note that C_2^d is the space of operators on $\mathbb{R}^d$ with the Hilbert-Schmidt norm. Also, for $p = \infty$, C_p^d is the standard ℓ_2 operator norm, and it is easy to verify that for $p = \log d$, $\|T\|_{2\to 2} \leq \|T\|_{C_p^d} \leq e\|T\|_{2\to 2}$.

The following inequality plays a central role in our analysis and is due to Lust-Piquard (see [9] for an exposition of that, and other results of a similar flavor). The estimate on the constant B_p was established by Rudelson [12].

Theorem 3. *There exists an absolute constant C, and for every $2 \leq p < \infty$ there is a constant B_p depending only on p, which satisfies $B_p \leq C\sqrt{p}$, for which the following holds. Let $y_1, ..., y_N$ be operators on $\mathbb{R}^d$, and denote*

$$A = \max\left\{ \left\| \left(\sum_{i=1}^N y_i^* y_i \right)^{1/2} \right\|_{C_p^d}, \left\| \left(\sum_{i=1}^N y_i y_i^* \right)^{1/2} \right\|_{C_p^d} \right\}.$$

Then,

$$A \leq \left(\mathbb{E}_\varepsilon \left\| \sum_{i=1}^N \varepsilon_i y_i \right\|_{C_p^d}^p \right)^{1/p} \leq B_p A.$$

We will use Theorem 3 for $y_i = x_i \otimes x_i$, and, as in Remark 1, without loss of generality, $y_i \in C_p^d$, for $d = \min\{n, N\}$. One can verify that in this case, $A = \left\| \left(\sum_{i=1}^N \|x_i\|^2 x_i \otimes x_i \right)^{1/2} \right\|_{C_p^d}$, and thus, for $p \geq 2$,

$$A \leq \left(\mathbb{E} \left\| \sum_{i=1}^N \varepsilon_i x_i \otimes x_i \right\|_{C_p^d}^p \right)^{1/p} \leq C\sqrt{p} A \tag{1.4}$$

The final preliminary result we require is Lidskii's inequality, on the differences of the sequences of the singular values of symmetric operators. For an operator T, denote by $\mu(T)$ the vector of singular values of T, arranged in a non-increasing order. Recall that for a vector $v \in \mathbb{R}^d$, and $1 \leq p < \infty$, $\|v\|_{\ell_p^d} = \left(\sum_{i=1}^d |v_i|^p \right)^{1/p}$ and for $p = \infty$, $\|v\|_\infty = \max_{1 \leq i \leq d} |v_i|$ (with obvious analogs for $d = \infty$).

Theorem 4. *[7] Let A and B be symmetric operators on $\mathbb{R}^d$. Then, for every $1 \leq p \leq \infty$, $\|\mu(A) - \mu(B)\|_{\ell_p^d} \leq \|\mu(A - B)\|_{\ell_p^d}$.*

The two most interesting cases here are $p = 2$ and $p = \infty$. For $p = 2$ it follows that the Euclidean distance between the vectors $\mu(A)$ and $\mu(B)$ is bounded by the Hilbert-Schmidt norm of $A-B$. For $p = \infty$, $\|\mu(A)-\mu(B)\|_{\ell_\infty^d} \leq \|A-B\|_{2\to 2}$.

2 Results

Let us begin with two estimates on the singular values of $\Gamma_N : \ell_2^N \to \ell_2$ defined by $\Gamma_N e_i = \frac{1}{\sqrt{N}} X_i$. Clearly, the nonzero eigenvalues of $\Gamma_N \Gamma_N^* = \frac{1}{N}\sum_{i=1}^N X_i \otimes X_i$, denoted by $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ...$, are the same as the nonzero eigenvalues of the Gram matrix $\left(\frac{1}{N}\langle X_i, X_j\rangle\right)_{i,j=1}^N$. As a notational convention, we will extend the finite vector $(\hat{\lambda}_1, ..., \hat{\lambda}_N)$ to an infinite one, by adding 0 in the $N+1$ component and beyond. Thus, one can consider the ℓ_2 and ℓ_∞ norms of the difference $\lambda - \hat{\lambda}$.

Our aim is to compare the eigenvalues of $\Gamma_N \Gamma_N^*$ to those of the average operator $\mathbb{E}(X \otimes X)$ (denoted by $\lambda_1 \geq \lambda_2 \geq ...$) with respect to the two norms. Since both $\sum_{i=1}^N X_i \otimes X_i$ and $\mathbb{E}(X \otimes X)$ are symmetric, and as long as $\mathbb{E}(X \otimes X)$ is in the appropriate Schatten class, then by Theorem 4 and approximating $\mathbb{E}(X \otimes X)$ by a finite dimensional operator, it follows that

$$\|\lambda - \hat{\lambda}\|_\infty = \sup_i \left|\lambda_i - \hat{\lambda}_i\right| \leq \left\| \frac{1}{N}\sum_{i=1}^N X_i \otimes X_i - \mathbb{E}(X \otimes X) \right\|_{2\to 2},$$

$$\|\lambda - \hat{\lambda}\|_2 = \left(\sum_{i=1}^\infty |\lambda_i - \hat{\lambda}_i|^2\right)^{1/2} \leq \left\| \frac{1}{N}\sum_{i=1}^N X_i \otimes X_i - \mathbb{E}(X \otimes X) \right\|_{C_2}.$$

The following bounds the expectation of the two norms of $\frac{1}{N}\sum_{i=1}^N X_i \otimes X_i - \mathbb{E}(X \otimes X)$. Its first part is a minor extension to a result due to Rudelson [12].

Theorem 5. *There exists an absolute constant C for which the following holds. Let X be a random vector in ℓ_2^n (resp. ℓ_2), set $d = \min\{N, n\}$ put $Q_N = (\mathbb{E}\max_{1\leq i\leq N}\|X_i\|^2)^{1/2}$, recall that $\Lambda = \mathbb{E}(X \otimes X)$ and set $T = \frac{1}{N}\sum_{i=1}^N X_i \otimes X_i - \Lambda$. Then,*

$$\mathbb{E}\,\|T\|_{2\to 2} \leq C \max\left\{ \frac{\log d}{N} Q_N^2,\ \min\left\{ \|\Lambda\|_{2\to 2}, \sqrt{\frac{\log d}{N}} \|\Lambda\|_{2\to 2}^{1/2} Q_N \right\} \right\}.$$

Also, if $\mathbb{E}\|X\|^4 < \infty$, then, $\mathbb{E}\,\|T\|_{C_2} \leq \frac{C}{\sqrt{N}} \left(\mathbb{E}\|X\|^4\right)^{1/2}$.

Remark 2. *It follows from a standard integration argument (see, e.g. [14]) that if Z is a random variable with a bounded ψ_α norm, and if $Z_1, ..., Z_N$ are independent copies of Z then*

$$\left\| \max_{1\leq i\leq N} Z_i \right\|_{\psi_\alpha} \leq C\|Z\|_{\psi_\alpha} \log^{1/\alpha} N$$

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

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

In particular, if $Z=\|X\|$ has a bounded ψ_α norm, then one can bound Q_N using $\|Z\|_{\psi_\alpha}$.

Proof of Theorem 5. Because the first part of the claim is an easy extension of a result from [12] we omit its proof. Some of the ideas required are also used in the proof of Theorem 7, below.

Turning to the second part of the claim, using a symmetrization argument, Hölder's inequality and applying Theorem 3 for $Y_i = X_i\otimes X_i$, it follows that

$$\begin{aligned}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^n X_i\otimes X_i - \mathbb{E}(X\otimes X)\right\|_{C_2} &\le \frac{1}{N}\mathbb{E}_X\left(\mathbb{E}_\varepsilon\left\|\sum_{i=1}^N \varepsilon_i X_i\otimes X_i\right\|_{C_2}^2\right)^{1/2}\\ &\le \frac{C}{N}\mathbb{E}_X\left(\left\|\left(\sum_{i=1}^N \|X_i\|^2 X_i\otimes X_i\right)^{1/2}\right\|_{C_2}\right).\end{aligned}$$

Let $U_i=\|X_i\|X_i$ and set $(\hat{\mu}_i)_{i=1}^N$ to be the singular values of the symmetric operator $\sum_{i=1}^N U_i\otimes U_i$. Since the nonzero singular values of $\sum_{i=1}^N U_i\otimes U_i$ are the same as that of $(\langle U_i,U_j\rangle)_{i,j=1}^N$, then $\sum_{i=1}^N\hat{\mu}_i=\sum_{i=1}^N\|U_i\|^2=\sum_{i=1}^N\|X_i\|^4$. Hence,

$$\left\|\left(\sum_{i=1}^N U_i\otimes U_i\right)^{1/2}\right\|_{C_2} = \left(\sum_{i=1}^N\hat{\mu}_i\right)^{1/2} = \left(\sum_{i=1}^N\|X_i\|^4\right)^{1/2},$$

from which the claim follows. ∎

It is possible to obtain estimates (which are probably suboptimal) on higher moments of $\left\|\frac{1}{N}\sum_{i=1}^N X_i\otimes X_i-\mathbb{E}(X\otimes X)\right\|_{2\to 2}$, and thus establish a deviation inequality, even when $\|X\|$ is not bounded. Of course, if $\|X\|$ is a bounded variable, one can apply Talagrand's concentration inequality for uniformly bounded empirical processes. To prove the desired deviation inequality in the unbounded case, one uses a "high moment" analog of the first part of Theorem 5, which builds on Theorem 3 and on Rudelson's approach from [12].

Theorem 6. *There exists an absolute constant c such that for any integers n and N, any $x_1,\ldots,x_N\in\mathbb{R}^n$ and any $p\ge 1$,*

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

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

Note that this moment inequality immediately leads to a ψ_2 estimate on the random variable $\left\|\sum_{i=1}^N \varepsilon_i x_i \otimes x_i\right\|_{2\to 2}$.

Corollary 2. *There exists an absolute constant c such that for any integers n and N, any $x_1,...,x_N \in \mathbb{R}^n$ and any $t>0$,*

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

where $\Delta = c\sqrt{\log d}\left\|\sum_{i=1}^N x_i\otimes x_i\right\|_{2\to 2}^{1/2}\max_{1\leq i\leq N}\|x_i\|$ and $d=\min\{N,n\}$.

Let us formulate and prove the desired tail estimate.

Theorem 7. *There exists an absolute constant c for which the following holds. Let X be a random vector in ℓ_2^n (resp. ℓ_2) which satisfies Assumption 1 and set $Z=\|X\|$, $\Lambda = \mathbb{E}(X\otimes X)$ and $\beta=(1+2/\alpha)^{-1}$. For any integers n and N let $d=\min\{N,n\}$,*

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

Then, for $1\leq p<\infty$,

$$\left(\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^N (X_i\otimes X_i)-\Lambda\right\|_{2\to 2}^p\right)^{1/p}\leq cp^{\frac{1}{\beta}}\max\left\{\frac{\rho^2}{\sqrt{N}}+\|\Lambda\|_{2\to 2}^{1/2}A_{d,N},A_{d,N}^2\right\},$$

and thus,

$$\left(\mathbb{E}\left(\sup_i|\hat{\lambda}_i-\lambda_i|\right)^p\right)^{1/p}\leq cp^{\frac{1}{\beta}}\max\left\{\frac{\rho^2}{\sqrt{N}}+\lambda_1^{1/2}A_{d,N},A_{d,N}^2\right\}.$$

In particular, for any $x>0$

$$Pr\left(\left\|\sum_{i=1}^N (X_i\otimes X_i)-\Lambda\right\|_{2\to 2}\geq xN\right)\leq\exp\left(-\left(\frac{cx}{\max\{B_{d,N},A_{d,N}^2\}}\right)^\beta\right),$$

and the same tail estimate holds for $\sup_i\left|\lambda_i-\hat{\lambda}_i\right|$.

Proof. Consider the random variables

$$S=\left\|\frac{1}{N}\sum_{i=1}^N\varepsilon_i X_i\otimes X_i\right\|_{2\to 2} \quad\text{and}\quad V=\left\|\frac{1}{N}\sum_{i=1}^N(X_i\otimes X_i-\Lambda)\right\|_{2\to 2}.$$

It follows from Corollaries 1 and 2 that for any $t\geq c\sqrt{\rho^4/N}$,

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

where $\Delta = c\sqrt{\log d}\left\|\sum_{i=1}^N X_i \otimes X_i\right\|_{2\to 2}^{1/2} \max_{1\le i\le N}\|X_i\|$ for some absolute constant c. Setting c_0 to be the constant from Corollary 1, then by Fubini's Theorem and dividing the region of integration to $t \le c_0\sqrt{\rho^4/N}$ (in this range one has no control on $Pr(V \ge t)$) and $t > c_0\sqrt{\rho^4/N}$, it is evident that

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

for some new absolute constant c.

The second term is bounded by

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

for some new absolute constant c. Hence, setting $Z = \|X\|$ and applying Assumption 1 and (2.1), we arrive at

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

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

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

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

Therefore, $\|V\|_{\psi_\beta} \le c\max\left\{B_{d,N}, A_{d,N}^2\right\}$, from which the estimate of the Theorem follows by a standard argument. ∎

Let us give an example of how the previous theorem can be used to compare the spectrum of the integral operator T_K with that of the Gram matrix.

Corollary 3. *Let $K : \Omega \times \Omega \to \mathbb{R}$ be a positive definite kernel, such that $\|K\|_{\psi_\alpha} < \infty$ for some $\alpha \geq 1$. If (λ_i) is the spectrum of the integral operator (arranged in a non-increasing order) and $(\hat{\lambda}_i)$ is the spectrum of the Gram matrix $\left(\frac{1}{N}K(t_i,t_j)\right)_{i,j=1}^N$ also arranged in a non-increasing order, then*

1. *For $1 \leq p < \infty$,*

$$\left(\mathbb{E}\sup_i \left|\lambda_i - \hat{\lambda}_i\right|^p\right)^{1/p} \leq$$
$$cp^{1+2/\alpha}\|K(t,t)\|_{\psi_{\alpha/2}} \max\left\{\frac{\sqrt{\log d}\log^{1/\alpha} N}{\sqrt{N}}, \frac{\log d \log^{2/\alpha} N}{N}\right\}.$$

2. *If $\mathbb{E}K^2(t,t) < \infty$ then $\mathbb{E}\|\lambda - \hat{\lambda}\|_2 \leq C\left(\frac{\mathbb{E}K^2(t,t)}{N}\right)^{1/2}$.*

Note that the second part of Corollary 3 generalizes and improves the following Lemma (Lemma 4.1) from [8].

Lemma 1. *Let $K(x,y) = \sum_{i=1}^R \lambda_i \phi_i(x)\phi_i(y)$ for $R < \infty$, and set $\xi^2(R) = \sum_{1\leq i,j\leq R}(\lambda_i^2 + \lambda_j^2)\mathbb{E}\phi_i^2\phi_j^2$. Then,*

$$\mathbb{E}\delta_2^2(\lambda,\hat{\lambda}) \leq \frac{\xi^2(R)}{N} - 2\frac{\sum_{i=1}^R \lambda_i^2}{N},$$

where for $u, v \in \ell_2$ $\delta_2(u,v) = \inf_\pi \|u - \pi(v)\|_{\ell_2}$, π is a permutation of $\{1,...\}$ and $\pi(v) = (v_{\pi(1)}, ...)$.

Our result extends this lemma in several ways. First of all, ours is an infinite dimensional result. Second, for every finite dimensional kernel, $\xi^2(R) \geq \mathbb{E}K^2(t,t)$, and finally, $\delta_2(\lambda,\hat{\lambda}) \leq \|\lambda - \hat{\lambda}\|_2$.

Corollary 3 is different from the results in [15], where the difference between the empirical trace and the actual one, and between the "tails" of the traces $\sum_{d+1}^\infty \lambda_i$ and $\sum_{d+1}^\infty \hat{\lambda}_i$ were established, rather than the ℓ_∞ and ℓ_2 distances of the vectors of the singular values, as in Corollary 3.

2.1 Approximation by Ellipsoids

Turning to (2), we will see how, for a finite dimensional vector X, the random operator Γ_N (defined by $\Gamma_N e_i = \frac{1}{\sqrt{N}}X_i$) approximates the polar of the Binet ellipsoid (the latter is generated by the covariance structure of X and was defined in (1.1)). Such an approach could be helpful in the analysis of Question 2. Indeed, if $\Gamma_N B_2^N$ asymptotically converges to $\mathcal{E}_B^\circ$, then its principal directions must converge to the principal directions of $\mathcal{E}_B^\circ$, and thus, the best d-approximating subspace of $\Gamma_N B_2^N$ will coincide in the limit with the space spanned by the d largest principal directions of $\mathcal{E}_B^\circ$.

Fix an integer n, let X be a random vector in ℓ_2^n, set $\mathcal{E}_B$ to be the Binet ellipsoid generated by X, and put (ψ_i) to be the orthonormal basis of the

principal directions of $\mathcal{E}_B$. Without loss of generality, assume that $\mathcal{E}_B$ has full rank. Then, $X = \sum_{i=1}^n \langle X, \psi_i \rangle \psi_i$, the covariance operator can be represented in the basis (ψ_i) by the matrix $A = \mathrm{diag}(\lambda_1, ..., \lambda_n)$, and it is standard to verify that $\mathcal{E}_B = A^{-1/2} B_2^n$. Set $Y = A^{-1/2} X$, and observe that Y is an isotropic vector; that is, for every $y \in \ell_2^n$, $\mathbb{E} |\langle Y, y \rangle|^2 = \|y\|_{\ell_2^n}^2$. The question of how well $\mathcal{K} = \left\{ \sum_{i=1}^n a_i Y_i \ : \ \sum_{i=1}^n a_i^2 \leq 1 \right\}$ approximates a multiple of the Euclidean ball has been thoroughly studied (see, e.g., [10, 6, 2, 12, 3]) under various assumptions on the vector Y. To that end, one has to show that for every $y \in B_2^n$,

$$\left| \frac{1}{N} \sum_{i=1}^N \langle Y_i, y \rangle^2 - 1 \right| \leq \delta. \tag{2.2}$$

By duality, (2.2) is equivalent to

$$(1 - \delta') \sqrt{N} B_2^n \subset \mathcal{K} \subset (1 + \delta') \sqrt{N} B_2^n$$

for a suitable δ', that is, to

$$(1 - \delta') A^{1/2} B_2^n \subset \Gamma_N B_2^n \subset (1 + \delta') A^{1/2} B_2^n,$$

implying that, $\Gamma_N B_2^n$ is equivalent to the dual of the Binet ellipsoid. One can verify that $\sup_{\{y : \|y\| = 1\}} \left| \frac{1}{N} \sum_{i=1}^N \langle Y, y \rangle^2 - 1 \right| \leq \delta$ if and only if all the singular values of the random operator $e_i \to Y_i / \sqrt{N}$ satisfy $|\mu_i - 1| \leq \delta$. Therefore, it suffices to show that, with high probability,

$$\left\| \frac{1}{N} \sum_{i=1}^N Y_i \otimes Y_i - \mathrm{Id} \right\|_{2 \to 2} \leq \delta,$$

which is the question we studied in the previous section.

Note that to apply Theorem 5, it suffices to control the decay of the ℓ_2^n norm of the random vector $A^{-1/2} X$, which is

$$\left\| A^{-1/2} X \right\|_2^2 = \sum_{j=1}^n \frac{1}{\lambda_j} \langle X_i, \psi_j \rangle^2 = \sum_{j=1}^n \frac{\langle X_i, \psi_j \rangle^2}{\mathbb{E} |\langle X, \psi_j \rangle|^2}.$$

Define $f_j = \frac{\langle X, \psi_j \rangle}{(\mathbb{E} |\langle X, \psi_j \rangle|^2)^{1/2}}$, observe that $(f_i)_{i=1}^n$ are orthonormal with respect to $L_2(\mu)$ and that $\left\| A^{-1/2} X \right\|_2^2 = \sum_j f_j^2$.

The next corollary can be derived from Theorem 5.

Corollary 4. *There exists an absolute constant c for which the following holds. Let Y be an isotropic random vector in ℓ_2^n, put $Y_1, ..., Y_N$ to be independent copies of Y and set $Q_N = \left(\mathbb{E} \max_{1 \leq i \leq N} \|Y_i\|_2^2 \right)^{1/2}$. If $\frac{Q_N^2 \log n}{N} \leq 1$ then*

$$\mathbb{E} \left\| \frac{1}{N} \sum_{i=1}^N Y_i \otimes Y_i - \mathrm{Id} \right\|_{2 \to 2} \leq c \cdot Q_N \sqrt{\frac{\log n}{N}}.$$

From the corollary applied to the random vector $Y = A^{-1/2}X$, it follows that if $Z = \|Y\|_2$ and $\|Z\|_{\psi_2} \leq c\sqrt{n}$ then for $\delta = c \cdot Q_N \sqrt{\log n/N} \leq c\sqrt{n/N}\log n$, with high probability,

$$(1-\delta)\mathcal{E}_B^\circ \subset \Gamma_N B_2^N \subset (1+\delta)\mathcal{E}_B^\circ.$$

Example. Let K be a finite dimensional, continuous kernel and set (ϕ_i) to be its Mercer basis. Then, $\langle X(t), \phi_i\rangle = \sqrt{\lambda_i}\phi_i(t)$, $f_i(t) = \phi_i(t)$, and $Y = (\phi_i(t))_{i=1}^n$. Assume that the eigenfunctions are bounded by M (such a bound always exists because each eigenfunction is bounded by Mercer's Theorem, and there is a finite number of eigenfunctions). Thus , $Z \equiv \|Y\|_2 \leq M\sqrt{n}$ and the same holds for Q_N. Therefore, if $N \geq c(M)n\log n$, $\Gamma_N B_2^N$ is a good approximation of the ellipsoid $\left\{\sum_{i=1}^n a_i\sqrt{\lambda_i}\phi_i \ : \ \sum_{i=1}^n a_i^2 \leq 1\right\}$.

2.2 Some Remarks

The assumption that $\|Z\|_{\psi_\alpha} \leq c\sqrt{n}$ is the best that one can hope for. Indeed, $\|Z\|_{\psi_\alpha} \geq c_\alpha \left(\mathbb{E}Z^2\right)^{1/2} \geq c_\alpha\sqrt{n}$. It also says something about the geometry of the random vector, since it implies that it is impossible for many of the functions f_i to be "peaked" at the same place. The most extreme case in which this condition holds is when the functions $\langle X, \psi_i\rangle$ are supported on disjoint sets of measure $1/n$, which implies that X is always in the direction of one of the ψ_is. More generally, the condition means that the random vector X can not have a components "much larger" than $\sqrt{\lambda_j}$ in many of the directions ψ_j simultaneously. For example, if $A = \{i : |\langle X, \psi_j\rangle| \geq \sqrt{t\lambda_j}\}$, then by the ψ_α assumption,

$$Pr\left(\{|A| \geq k\}\right) \leq Pr\left(\left\{\sum_{i=1}^n f_j^2 \geq kt\right\}\right) \leq 2\exp\left(-c\left(\frac{kt}{n}\right)^{\alpha/2}\right).$$

Let us mention that such an assumption on the random vector is not that far-fetched. First of all, if $\mathcal{E}$ is an n dimensional ellipsoid in $L_2(\mu)$, one can find orthonormal vectors ϕ_i and positive scalars θ_i, such that

$$\mathcal{E} = \left\{\sum_{i=1}^n a_i\sqrt{\theta_i}\phi_i \ : \ \sum_{i=1}^n a_i^2 \leq 1\right\}$$

and $\sum_{i=1}^n \phi_i^2 = n$ pointwise. This basis is a simple example of the so-called *Lewis basis*, which has many applications in convex geometry (see, for example, [5]). Hence, one can approximate any ellipsoid by the random ellipsoid $\Gamma_N B_2^N$ using $X(t) = \sum_{i=1}^n \sqrt{\theta_i}\phi_i(t)\phi_i$.

The second remark we wish to make is that if Y is an isotropic vector in $\mathbb{R}^n$ which distributed according to a log-concave measure, and if $Z = \|Y\|$, then $\|Z\|_{\psi_2} \leq c\sqrt{n}$. This fact was shown in [3], and generalized the analogous result for a random point selected from a convex body, due to Alesker [1].

To conclude, because this notion of approximation is very strong, one must impose restrictive conditions on the random vector X which also depend on

the structure of the eigenfunctions. Perhaps a possible way of improving the rate of $\sqrt{n/N}\log n$ is to consider a weaker notion of approximation, namely that $\hat{\mathcal{E}} \subset \mathcal{E} + \delta B_2^n$ and $\mathcal{E} \subset \hat{\mathcal{E}} + \delta B_2^n$. It seems likely that for this notion of approximation, one could use the fact that $\mathcal{E}$ has small eigenvalues and obtain a better bound. The disadvantage is that the analysis of this question could be difficult, because one has to simultaneously control three different Euclidean structures (of $\mathcal{E}$, $\hat{\mathcal{E}}$ and B_2^n), and thus we leave it open for further investigation.

References

1. S. Alesker, ψ_2 estimates for the Euclidean norm on a convex body in isotropic position, Operator Theory Adv. Appl. 77, 1-4 1995.
2. J. Bourgain, Random points in isotropic convex bodies, in Convex Geometric Analysis (Berkeley, CA, 1996) Math. Sci. Res. Inst. Publ. 34 (1999), 53-58.
3. A.A. Giannopoulos, V.D. Milman, Concentration property on probability spaces, Adv. Math. 156, 77-106, 2000.
4. R. Herbrich, *Learning kernel classifiers*, MIT Press, 2002.
5. W.B. Johnson, G. Schechtman: Finite dimensional subspaces of L_p, in *Handbook of the Geometry of Banach Spaces, Vol 1* (W.B. Johnson, J. Lindenstrauss eds.), North Holland, 2001.
6. R. Kannan, L. Lovász, M. Simonovits, Random walks and $O^*(n^5)$ volume algorithm for convex bodies, Random structures and algorithms, 2(1) 1-50, 1997.
7. T. Kato, *A short introduction to perturbation theory for linear operators*, Springer-Verlag, 1982.
8. V. Koltchinskii, E. Giné, Random matrix approximation of spectra of integral operators. Bernoulli, 6 (2000) 113-167.
9. F. Lust-Piquard, G. Pisier, Non-commutative Khinchine and Paley inequalities, Ark. Mat. 29, 241-260, 1991.
10. V.D. Milman, A. Pajor, Isotropic position and inertia ellipsoids and zonoids of the unit ball of a normed n-dimensional space, Lecture notes in mathematics 1376, 64-104, Springer, 1989.
11. B. Schölkopf, A.J. Smola, *Learning with kernels*, MIT Press, 2002.
12. M. Rudelson, Random vectors in the isotropic position, Journal of Functional Analysis, 164, 60-72, 1999.
13. M. Talagrand, *The generic chaining*, forthcoming.
14. A.W. Van der Vaart, J.A. Wellner, *Weak convergence and Empirical Processes*, Springer-Verlag, 1996.
15. L. Zwald, O. Bousquet, G. Blanchard, Statistical properties of kernel principal component analysis, in *Proceedings of COLT 2004, J. Shawe-Taylor and Y. Singer (Eds)*, LNAI 3120, 594-608, Springer-Verlag 2004.