Approximation properties of certain operator-induced norms on Hilbert spaces

Arash A. Amini[b], Martin J. Wainwright[a,b]

[a]*Department of Statistics and*
[b]*Department of Electrical Engineering and Computer Sciences UC Berkeley, Berkeley, CA 94720*

Abstract

We consider a class of operator-induced norms, acting as finite-dimensional surrogates to the L^2 norm, and study their approximation properties over Hilbert subspaces of L^2. The class includes, as a special case, the usual empirical norm encountered, for example, in the context of nonparametric regression in reproducing kernel Hilbert spaces (RKHS). Our results have implications to the analysis of M-estimators in models based on finite-dimensional linear approximation of functions, and also to some related packing problems.

Keywords:
L^2 approximation, Empirical norm, Quadratic functionals, Hilbert spaces with reproducing kernels, Analysis of M-estimators

1. Introduction

Given a probability measure $\mathbb{P}$ supported on a compact set $\mathcal{X} \subset \mathbb{R}^d$, consider the function class

$$L^2(\mathbb{P}) := \big\{ f : \mathcal{X} \to \mathbb{R} \;|\; \|f\|_{L^2(\mathbb{P})} < \infty \big\}, \tag{1}$$

where $\|f\|_{L^2(\mathbb{P})} := \sqrt{\int_{\mathcal{X}} f^2(x)\, d\mathbb{P}(x)}$ is the usual L^2 norm[1] defined with respect to the measure $\mathbb{P}$. It is often of interest to construct approximations

Email addresses: `amini@eecs.berkeley.edu` (Arash A. Amini), `wainwrig@stat.berkeley.edu` (Martin J. Wainwright)

[1]We also use $L^2(\mathcal{X})$ or simply L^2 to refer to the space (1), with corresponding conventions for its norm. Also, one can take $\mathcal{X}$ to be a compact subset of any separable metric space and $\mathbb{P}$ a (regular) Borel measure.

to this L^2 norm that are "finite-dimensional" in nature, and to study the quality of approximation over the unit ball of some Hilbert space $\mathcal{H}$ that is continuously embedded within L^2. For example, in approximation theory and mathematical statistics, a collection of n design points in $\mathcal{X}$ is often used to define a surrogate for the L^2 norm. In other settings, one is given some orthonormal basis of $L^2(\mathbb{P})$, and defines an approximation based on the sum of squares of the first n (generalized) Fourier coefficients. For problems of this type, it is of interest to gain a precise understanding of the approximation accuracy in terms of its dimension n and other problem parameters.

The goal of this paper is to study such questions in reasonable generality for the case of Hilbert spaces $\mathcal{H}$. We let $\Phi_n : \mathcal{H} \to \mathbb{R}^n$ denote a continuous linear operator on the Hilbert space, which acts by mapping any $f \in \mathcal{H}$ to the n-vector $\big([\Phi_n f]_1 \quad [\Phi_n f]_2 \quad \cdots \quad [\Phi_n f]_n\big)$. This operator defines the Φ_n-semi-norm

$$\|f\|_{\Phi_n} := \sqrt{\sum_{i=1}^{n} [\Phi_n f]_i^2}. \tag{2}$$

In the sequel, with a minor abuse of terminology,[2] we refer to $\|f\|_{\Phi_n}$ as the Φ_n-norm of f. Our goal is to study how well $\|f\|_{\Phi_n}$ approximates $\|f\|_{L^2}$ over the unit ball of $\mathcal{H}$ as a function of n, and other problem parameters. We provide a number of examples of the *sampling operator* Φ_n in Section 2.2. Since the dependence on the parameter n should be clear, we frequently omit the subscript to simplify notation.

In order to measure the quality of approximation over $\mathcal{H}$, we consider the quantity

$$R_\Phi(\varepsilon) := \sup\big\{\|f\|_{L^2}^2 \;\big|\; f \in B_{\mathcal{H}},\; \|f\|_\Phi^2 \;\le \varepsilon^2\big\}, \tag{3}$$

where $B_{\mathcal{H}} := \{f \in \mathcal{H} \mid \|f\|_{\mathcal{H}} \le 1\}$ is the unit ball of $\mathcal{H}$. The goal of this paper is to obtain sharp upper bounds on R_Φ. As discussed in Appendix Appendix C, a relatively straightforward argument can be used to translate such upper bounds into lower bounds on the related quantity

$$\underline{T}_\Phi(\varepsilon) := \inf\; \big\{\|f\|_\Phi^2 \;\big|\; f \in B_{\mathcal{H}},\; \|f\|_{L^2}^2 \ge \varepsilon^2\big\}. \tag{4}$$

[2]This can be justified by identifying f and g if $\Phi f = \Phi g$, i.e. considering the quotient $\mathcal{H}/\ker\Phi$.

We also note that, for a complete picture of the relationship between the semi-norm $\|\cdot\|_\Phi$ and the L^2 norm, one can also consider the related pair

$$T_\Phi(\varepsilon) := \sup\big\{\|f\|_\Phi^2 \;\mid\; f \in B_{\mathcal{H}},\; \|f\|_{L^2}^2 \le \varepsilon^2\big\}, \quad \text{and} \tag{5a}$$

$$\underline{R}_\Phi(\varepsilon) := \inf\big\{\|f\|_{L^2}^2 \;\mid\; f \in B_{\mathcal{H}},\; \|f\|_\Phi^2 \ge \varepsilon^2\big\}. \tag{5b}$$

Our methods are also applicable to these quantities, but we limit our treatment to $(R_\Phi, \underline{T}_\Phi)$ so as to keep the contribution focused.

Certain special cases of linear operators Φ, and associated functionals have been studied in past work. In the special case $\varepsilon = 0$, we have

$$R_\Phi(0) = \sup\big\{\|f\|_{L^2}^2 \;\mid\; f \in B_{\mathcal{H}},\; \Phi(f) = 0\big\},$$

a quantity that corresponds to the squared diameter of $B_{\mathcal{H}} \cap \mathrm{Ker}(\Phi)$, measured in the L^2-norm. Quantities of this type are standard in approximation theory (e.g., [1, 2, 3]), for instance in the context of Kolmogorov and Gelfand widths. Our primary interest in this paper is the more general setting with $\varepsilon > 0$, for which additional factors are involved in controlling $R_\Phi(\varepsilon)$. In statistics, there is a literature on the case in which Φ is a sampling operator, which maps each function f to a vector of n samples, and the norm $\|\cdot\|_\Phi$ corresponds to the empirical L^2-norm defined by these samples. When these samples are chosen randomly, then techniques from empirical process theory [4] can be used to relate the two terms. As discussed in the sequel, our results have consequences for this setting of random sampling.

As an example of a problem in which an upper bound on R_Φ is useful, let us consider a general linear inverse problem, in which the goal is to recover an estimate of the function f^* based on the noisy observations

$$y_i = [\Phi f^*]_i + w_i, \quad i = 1, \ldots, n,$$

where $\{w_i\}$ are zero-mean noise variables, and $f^* \in B_{\mathcal{H}}$ is unknown. An estimate $\widehat{f}$ can be obtained by solving a least-squares problem over the unit ball of the Hilbert space—that is, to solve the convex program

$$\widehat{f} := \arg\min_{f \in B_{\mathcal{H}}} \sum_{i=1}^{n} (y_i - [\Phi f]_i)^2.$$

For such estimators, there are fairly standard techniques for deriving upper bounds on the Φ-semi-norm of the deviation $\widehat{f} - f^*$. Our results in this paper on R_Φ can then be used to translate this to a corresponding upper bound on the L^2-norm of the deviation $\widehat{f} - f^*$, which is often a more natural measure of performance.

As an example where the dual quantity $\underline{T}_\Phi$ might be helpful, consider the packing problem for a subset $\mathcal{D} \subset B_{\mathcal{H}}$ of the Hilbert ball. Let $M(\varepsilon; \mathcal{D}, \|\cdot\|_{L^2})$ be the ε-packing number of $\mathcal{D}$ in $\|\cdot\|_{L^2}$, i.e., the maximal number of function $f_1, \dots, f_M \in \mathcal{D}$ such that $\|f_i - f_j\|_{L^2} \geq \varepsilon$ for all $i, j = 1, \dots, M$. Similarly, let $M(\varepsilon; \mathcal{D}, \|\cdot\|_\Phi)$ be the ε-packing number of $\mathcal{D}$ in $\|\cdot\|_\Phi$ norm. Now, suppose that for some fixed ε, $\underline{T}_\Phi(\varepsilon) > 0$. Then, if we have a collection of functions $\{f_1, \dots, f_M\}$ which is an ε-packing of $\mathcal{D}$ in $\|\cdot\|_{L^2}$ norm, then the same collection will be a $\sqrt{\underline{T}_\Phi(\varepsilon)}$-packing of $\mathcal{D}$ in $\|\cdot\|_\Phi$. This implies the following useful relationship between packing numbers

$$M(\varepsilon\,; \mathcal{D}, \|\cdot\|_{L^2}) \leq M(\sqrt{\underline{T}_\Phi(\varepsilon)}\,; \mathcal{D}, \|\cdot\|_\Phi).$$

The remainder of this paper is organized as follows. We begin in Section 2 with background on the Hilbert space set-up, and provide various examples of the linear operators Φ to which our results apply. Section 3 contains the statement of our main result, and illustration of some its consequences for different Hilbert spaces and linear operators. Finally, Section 4 is devoted to the proofs of our results.

Notation:. For any positive integer p, we use $\mathbb{S}_+^p$ to denote the cone of $p \times p$ positive semidefinite matrices. For $A, B \in \mathbb{S}_+^p$, we write $A \succeq B$ or $B \preceq A$ to mean $A - B \in \mathbb{S}_+^p$. For any square matrix A, let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote its minimal and maximal eigenvalues, respectively. We will use both $\sqrt{A}$ and $A^{1/2}$ to denote the symmetric square root of $A \in \mathbb{S}_+^p$. We will use $\{x_k\} = \{x_k\}_{k=1}^\infty$ to denote a (countable) sequence of objects (e.g. real-numbers and functions). Occasionally we might denote an n-vector as $\{x_1, \dots, x_n\}$. The context will determine whether the elements between braces are ordered. The symbols $\ell_2 = \ell_2(\mathbb{N})$ are used to denote the Hilbert sequence space consisting of real-valued sequences equipped with the inner product $\langle \{x_k\}, \{y_k\} \rangle_{\ell_2} := \sum_{k=1}^\infty x_i y_i$. The corresponding norm is denoted as $\|\cdot\|_{\ell_2}$.

2. Background

We begin with some background on the class of Hilbert spaces of interest in this paper and then proceed to provide some examples of the sampling operators of interest.

2.1. Hilbert spaces

We consider a class of Hilbert function spaces contained within $L^2(\mathcal{X})$, and defined as follows. Let $\{\psi_k\}_{k=1}^\infty$ be an orthonormal sequence (not necessarily a basis) in $L^2(\mathcal{X})$ and let $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \cdots > 0$ be a sequence of positive weights decreasing to zero. Given these two ingredients, we can consider the class of functions

$$\mathcal{H} := \Big\{ f \in L^2(\mathbb{P}) \;\Big|\; f = \sum_{k=1}^\infty \sqrt{\sigma_k}\alpha_k\psi_k, \quad \text{for some } \{\alpha_k\}_{k=1}^\infty \in \ell_2(\mathbb{N}) \Big\}, \tag{6}$$

where the series in (6) is assumed to converge in L^2. (The series converges since $\sum_{k=1}^\infty (\sqrt{\sigma_k}\alpha_k)^2 \leq \sigma_1 \|\{\alpha_k\}\|_{\ell_2} < \infty$.) We refer to the sequence $\{\alpha_k\}_{k=1}^\infty \in \ell_2$ as the representative of f. Note that this representation is unique due to σ_k being strictly positive for all $k \in \mathbb{N}$.

If f and g are two members of $\mathcal{H}$, say with associated representatives $\alpha = \{\alpha_k\}_{k=1}^\infty$ and $\beta = \{\beta_k\}_{k=1}^\infty$, then we can define the inner product

$$\langle f, g\rangle_{\mathcal{H}} := \sum_{k=1}^\infty \alpha_k\beta_k \;=\; \langle \alpha, \beta\rangle_{\ell_2}. \tag{7}$$

With this choice of inner product, it can be verified that the space $\mathcal{H}$ is a Hilbert space. (In fact, $\mathcal{H}$ inherits all the required properties directly from ℓ_2.) For future reference, we note that for two functions $f, g \in \mathcal{H}$ with associated representatives $\alpha, \beta \in \ell_2$, their L^2-based inner product is given by[3] $\langle f, g\rangle_{L^2} = \sum_{k=1}^\infty \sigma_k\alpha_k\beta_k$.

We note that each ψ_k is in $\mathcal{H}$, as it is represented by a sequence with a single nonzero element, namely, the k-th element which is equal to $\sigma_k^{-1/2}$. It follows from (7) that $\langle \sqrt{\sigma_k}\psi_k, \sqrt{\sigma_j}\psi_j\rangle_{\mathcal{H}} = \delta_{kj}$. That is, $\{\sqrt{\sigma_k}\psi_k\}$ is an orthonormal sequence in $\mathcal{H}$. Now, let $f \in \mathcal{H}$ be represented by $\alpha \in \ell_2$. We claim that the series in (6) also converges in $\mathcal{H}$ norm. In particular, $\sum_{k=1}^N \sqrt{\sigma_k}\alpha_k\psi_k$ is in $\mathcal{H}$, as it is represented by the sequence $\{\alpha_1, \dots, \alpha_N, 0, 0, \dots\} \in \ell_2$. It follows from (7) that $\|f - \sum_{k=1}^N \sqrt{\sigma_k}\alpha_k\psi_k\|_{\mathcal{H}} = \sum_{k=N+1}^\infty \alpha_k^2$ which converges to 0 as $N \to \infty$. Thus, $\{\sqrt{\sigma_k}\psi_k\}$ is in fact an orthonormal basis for $\mathcal{H}$.

[3]In particular, for $f \in \mathcal{H}$, $\|f\|_{L^2} \leq \sqrt{\sigma_1}\|f\|_{\mathcal{H}}$ which shows that the inclusion $\mathcal{H} \subset L^2$ is continuous.

We now turn to a special case of particular importance to us, namely the reproducing kernel Hilbert space (RKHS) of a continuous kernel. Consider a symmetric bivariate function $\mathbb{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, where $\mathcal{X} \subset \mathbb{R}^d$ is compact[4]. Furthermore, assume $\mathbb{K}$ to be positive semidefinite and continuous. Consider the integral operator $I_{\mathbb{K}}$ mapping a function $f \in L^2$ to the function $I_{\mathbb{K}} f := \int \mathbb{K}(\cdot, y) f(y) d\mathbb{P}(y)$. As a consequence of Mercer's theorem [5, 6], $I_{\mathbb{K}}$ is a compact operator from L^2 to $C(\mathcal{X})$, the space of continuous functions on $\mathcal{X}$ equipped with the uniform norm[5]. Let $\{\sigma_k\}$ be the sequence of nonzero eigenvalues of $I_{\mathbb{K}}$, which are positive, can be ordered in nonincreasing order and converge to zero. Let $\{\psi_k\}$ be the corresponding eigenfunctions which are continuous and can be taken to be orthonormal in L^2. With these ingredients, the space $\mathcal{H}$ defined in equation (6) is the RKHS of the kernel function $\mathbb{K}$. This can be verified as follows.

As another consequence of the Mercer's theorem, $\mathbb{K}$ has the decomposition

$$\mathbb{K}(x, y) := \sum_{k=1}^{\infty} \sigma_k \psi_k(x) \psi_k(y) \tag{8}$$

where the convergence is absolute and uniform (in x and y). In particular, for any fixed $y \in \mathcal{X}$, the sequence $\{\sqrt{\sigma_k}\psi_k(y)\}$ is in ℓ_2. (In fact, $\sum_{k=1}^{\infty} (\sqrt{\sigma_k}\psi_k(y))^2 = \mathbb{K}(y, y) < \infty$.) Hence, $\mathbb{K}(\cdot, y)$ is in $\mathcal{H}$, as defined in (6), with representative $\{\sqrt{\sigma_k}\psi_k(y)\}$. Furthermore, it can be verified that the convergence in (6) can be taken to be also pointwise[6]. To be more specific, for any $f \in \mathcal{H}$ with representative $\{\alpha_k\}_{k=1}^{\infty} \in \ell_2$, we have $f(y) = \sum_{k=1}^{\infty} \sqrt{\sigma_k}\alpha_k \psi_k(y)$, for all $y \in \mathcal{X}$. Consequently, by definition of the inner product (7), we have

$$\langle f, \mathbb{K}(\cdot, y) \rangle_{\mathcal{H}} = \sum_{k=1}^{\infty} \alpha_k \sqrt{\sigma_k} \psi_k(y) \; = \; f(y),$$

so that $\mathbb{K}(\cdot, y)$ acts as the representer of evaluation. This argument shows that for any fixed $y \in \mathcal{X}$, the linear functional on $\mathcal{H}$ given by $f \mapsto f(y)$ is

[4] Also assume that $\mathbb{P}$ assign positive mass to every open Borel subset of $\mathcal{X}$.

[5] In fact, $I_{\mathbb{K}}$ is well defined over $L^1 \supset L^2$ and the conclusions about $I_{\mathbb{K}}$ hold as a operator from L^1 to $C(\mathcal{X})$.

[6] The convergence is actually even stronger, namely it is absolute and uniform, as can be seen by noting that $\sum_{k=n+1}^{m} |\alpha_k \sqrt{\sigma_k} \psi_k(y)| \le (\sum_{k=n+1}^{m} \alpha_k^2)^{1/2} (\sum_{k=n+1}^{m} \sigma_k \psi_k^2(y))^{1/2} \le (\sum_{k=n+1}^{m} \alpha_k^2)^{1/2} \max_{y \in \mathcal{X}} k(y, y)$.

bounded, since we have

$$|f(y)| = \big|\langle f, \mathbb{K}(\cdot, y)\rangle_{\mathcal{H}}\big| \;\leq\; \|f\|_{\mathcal{H}} \|\mathbb{K}(\cdot, y)\|_{\mathcal{H}},$$

hence $\mathcal{H}$ is indeed the RKHS of the kernel $\mathbb{K}$. This fact plays an important role in the sequel, since some of the linear operators that we consider involve pointwise evaluation.

A comment regarding the scope: our general results hold for the basic setting introduced in equation (6). For those examples that involve pointwise evaluation, we assume the more refined case of the RKHS described above.

2.2. Linear operators, semi-norms and examples

Let $\Phi : \mathcal{H} \to \mathbb{R}^n$ be a continuous linear operator, with co-ordinates $[\Phi f]_i$ for $i = 1, 2, \ldots, n$. It defines the (semi)-inner product

$$\langle f, g\rangle_{\Phi} := \langle \Phi f, \Phi g\rangle_{\mathbb{R}^n}, \tag{9}$$

which induces the semi-norm $\|\cdot\|_{\Phi}$. By the Riesz representation theorem, for each $i = 1, \ldots, n$, there is a function $\varphi_i \in \mathcal{H}$ such that $[\Phi f]_i = \langle \varphi_i, f\rangle_{\mathcal{H}}$ for any $f \in \mathcal{H}$.

Let us illustrate the preceding definitions with some examples.

Example 1 (Generalized Fourier truncation)**.** Recall the orthonormal basis $\{\psi_i\}_{i=1}^{\infty}$ underlying the Hilbert space. Consider the linear operator $\mathbb{T}_{\psi_1^n} : \mathcal{H} \to \mathbb{R}^n$ with coordinates

$$[\mathbb{T}_{\psi_1^n} f]_i := \langle \psi_i, f\rangle_{L^2}, \quad \text{for } i = 1, 2, \ldots, n. \tag{10}$$

We refer to this operator as the *(generalized) Fourier truncation operator,* since it acts by truncating the (generalized) Fourier representation of f to its first n co-ordinates. More precisely, by construction, if $f = \sum_{k=1}^{\infty} \sqrt{\sigma_k}\alpha_k\psi_k$, then

$$[\Phi f]_i = \sqrt{\sigma_i}\alpha_i, \qquad \text{for } i = 1, 2, \ldots, n. \tag{11}$$

By definition of the Hilbert inner product, we have $\alpha_i = \langle \psi_i, f\rangle_{\mathcal{H}}$, so that we can write $[\Phi f]_i = \langle \varphi_i, f\rangle_{\mathcal{H}}$, where $\varphi_i := \sqrt{\sigma_i}\psi_i$. ♢

Example 2 (Domain sampling)**.** A collection $x_1^n := \{x_1, \ldots, x_n\}$ of points in the domain $\mathcal{X}$ can be used to define the (scaled) *sampling operator* $\mathbb{S}_{x_1^n} : \mathcal{H} \to \mathbb{R}^n$ via

$$\mathbb{S}_{x_1^n} f := n^{-1/2} \big(f(x_1) \quad \ldots \quad f(x_n) \big), \quad \text{for } f \in \mathcal{H}. \tag{12}$$

As previously discussed, when $\mathcal{H}$ is a reproducing kernel Hilbert space (with kernel $\mathbb{K}$), the (scaled) evaluation functional $f \mapsto n^{-1/2} f(x_i)$ is bounded, and its Riesz representation is given by the function $\varphi_i = n^{-1/2}\mathbb{K}(\cdot, x_i)$. ♢

Example 3 (Weighted domain sampling)**.** Consider the setting of the previous example. A slight variation on the sampling operator (12) is obtained by adding some weights to the samples

$$\mathbb{W}_{x_1^n, w_1^n} f := n^{-1/2} \big(w_1 f(x_1) \quad \ldots \quad w_n f(x_n) \big), \quad \text{for } f \in \mathcal{H}. \tag{13}$$

where $w_1^n = (w_1, \ldots, w_n)$ is chosen such that $\sum_{k=1}^n w_k^2 = 1$. Clearly, $\varphi_i = n^{-1/2} w_i \, \mathbb{K}(\cdot, x_i)$.

[As an example of how this might arise, consider approximating $f(t)$ by $\sum_{k=1}^n f(x_k) G_n(t, x_k)$ where $\{G_n(\cdot\,, x_k)\}$ is a collection of functions in $L^2(\mathcal{X})$ such that $\langle G_n(\cdot\,, x_k), G_n(\cdot\,, x_j)\rangle_{L^2} = n^{-1} w_k^2 \, \delta_{kj}$. Proper choices of $\{G_n(\cdot, x_i)\}$ might produce better approximations to the L^2 norm in the cases where one insists on choosing elements of x_1^n to be uniformly spaced, while $\mathbb{P}$ in (1) is not a uniform distribution. Another slightly different but closely related case is when one approximates $f^2(t)$ over $\mathcal{X} = [0,1]$, by say $n^{-1} \sum_{k=1}^{n-1} f^2(x_k) W(n(t - x_k))$ for some function $W : [-1, 1] \to \mathbb{R}_+$ and $x_k = k/n$. Again, non-uniform weights are obtained when $\mathbb{P}$ is nonuniform.]

♢

3. Main result and some consequences

We now turn to the statement of our main result, and the development of some its consequences for various models.

3.1. General upper bounds on $R_\Phi(\varepsilon)$

We now turn to upper bounds on $R_\Phi(\varepsilon)$ which was defined previously in (3). Our bounds are stated in terms of a real-valued function defined as follows: for matrices $D, M \in \mathbb{S}_+^p$,

$$\mathcal{L}(t, M, D) := \max\Big\{ \lambda_{\max}\big(D - t\sqrt{D}\, M \sqrt{D}\big),\ 0 \Big\}, \qquad \text{for } t \geq 0. \tag{14}$$

Here $\sqrt{D}$ denotes the matrix square root, valid for positive semidefinite matrices.

The upper bounds on $R_\Phi(\varepsilon)$ involve principal submatrices of certain infinite-dimensional matrices—or equivalently linear operators on $\ell_2(\mathbb{N})$—that we define here. Let Ψ be the infinite-dimensional matrix with entries

$$[\Psi]_{jk} := \langle \psi_j, \psi_k \rangle_\Phi, \quad \text{for } j, k = 1, 2, \ldots, \tag{15}$$

and let $\Sigma = \text{diag}\{\sigma_1, \sigma_2, \ldots,\}$ be a diagonal operator. For any $p = 1, 2, \ldots$, we use Ψ_p and $\Psi_{\widetilde{p}}$ to denote the principal submatrices of Ψ on rows and columns indexed by $\{1, 2, \ldots, p\}$ and $\{p+1, p+2, \ldots\}$, respectively. A similar notation will be used to denote submatrices of Σ.

Theorem 1. *For all $\varepsilon \geq 0$, we have:*

$$R_\Phi(\varepsilon) \;\leq\; \inf_{p \in \mathbb{N}} \inf_{t \geq 0} \Big\{ \mathcal{L}(t, \Psi_p, \Sigma_p) + t \Big(\varepsilon + \sqrt{\lambda_{\max}(\Sigma_{\widetilde{p}}^{1/2} \Psi_{\widetilde{p}} \Sigma_{\widetilde{p}}^{1/2})} \Big)^2 + \sigma_{p+1} \Big\}. \tag{16}$$

Moreover, for any $p \in \mathbb{N}$ such that $\lambda_{\min}(\Psi_p) > 0$, we have

$$R_\Phi(\varepsilon) \;\leq\; \Big(1 - \frac{\sigma_{p+1}}{\sigma_1}\Big) \frac{1}{\lambda_{\min}(\Psi_p)} \Big(\varepsilon + \sqrt{\lambda_{\max}(\Sigma_{\widetilde{p}}^{1/2} \Psi_{\widetilde{p}} \Sigma_{\widetilde{p}}^{1/2})} \Big)^2 + \sigma_{p+1}. \tag{17}$$

Remark (a):. These bounds cannot be improved in general. This is most easily seen in the special case $\varepsilon = 0$. Setting $p = n$, bound (17) implies that $R_\Phi(0) \leq \sigma_{n+1}$ whenever Ψ_n is strictly positive definite and $\Psi_{\widetilde{n}} = 0$. This bound is sharp in a "minimax sense", meaning that equality holds if we take the infimum over all bounded linear operators $\Phi : \mathcal{H} \to \mathbb{R}^n$. In particular, it is straightforward to show that

$$\inf_{\substack{\Phi:\, \mathcal{H} \to \mathbb{R}^n \\ \Phi \text{ surjective}}} R_\Phi(0) \;=\; \inf_{\substack{\Phi:\, \mathcal{H} \to \mathbb{R}^n \\ \Phi \text{ surjective}}} \sup_{f \in B_{\mathcal{H}}} \big\{ \|f\|_{L^2}^2 \;\mid\; \Phi f = 0 \big\} \;=\; \sigma_{n+1}, \tag{18}$$

and moreover, this infimum is in fact achieved by some linear operator. Such results are known from the general theory of n-widths for Hilbert spaces (e.g., see Chapter IV in Pinkus [2] and Chapter 3 of [7].)

In the more general setting of $\varepsilon > 0$, there are operators for which the bound (17) is met with equality. As a simple illustration, recall the (generalized) Fourier truncation operator $\mathbb{T}_{\psi_1^n}$ from Example 1. First, it can be

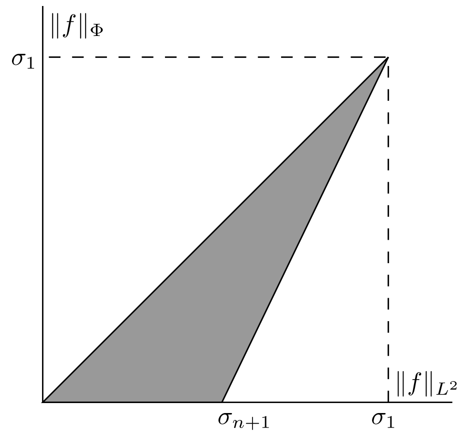


Figure 1: Geometry of Fourier truncation. The plot shows the set $\{(\|f\|_{L^2}, \|f\|_\Phi) : \|f\|_{\mathcal{H}} \leq 1\} \subset \mathbb{R}^2$ for the case of (generalized) Fourier truncation operator $\mathbb{T}_{\psi_1^n}$.

verified that $\langle \psi_k, \psi_j \rangle_{\mathbb{T}_{\psi_1^n}} = \delta_{jk}$ for $j, k \leq n$ and $\langle \psi_k, \psi_j \rangle_{\mathbb{T}_{\psi_1^n}} = 0$ otherwise. Taking $p = n$, we have $\Psi_n = I_n$, that is, the n-by-n identity matrix, and $\Psi_{\widetilde{n}} = 0$. Taking $p = n$ in (17), it follows that for $\varepsilon^2 \leq \sigma_1$,

$$R_{\mathbb{T}_{\psi_1^n}}(\varepsilon) \;\leq\; \Big(1 - \frac{\sigma_{n+1}}{\sigma_1}\Big)\varepsilon^2 + \sigma_{n+1}, \tag{19}$$

As shown in Appendix Appendix E, the bound (19) in fact holds with equality. In other words, the bounds of Theorems 1 are tight in this case. Also, note that (19) implies $R_{\mathbb{T}_{\psi_1^n}}(0) \leq \sigma_{n+1}$ showing that the (generalized) Fourier truncation operator achieves the minimax bound of (18). Fig 1 provides a geometric interpretation of these results.

Remark (b):. In general, it might be difficult to obtain a bound on $\lambda_{\max}(\Sigma_{\widetilde{p}}^{1/2}\Psi_{\widetilde{p}}\Sigma_{\widetilde{p}}^{1/2})$ as it involves the infinite dimensional matrix $\Psi_{\widetilde{p}}$. One may obtain a simple (although not usually sharp) bound on this quantity by noting that for a positive semidefinite matrix, the maximal eigenvalue is bounded by the trace, that is,

$$\lambda_{\max}\big(\Sigma_{\widetilde{p}}^{1/2}\Psi_{\widetilde{p}}\Sigma_{\widetilde{p}}^{1/2}\big) \leq \operatorname{tr}\big(\Sigma_{\widetilde{p}}^{1/2}\Psi_{\widetilde{p}}\Sigma_{\widetilde{p}}^{1/2}\big) = \sum_{k>p} \sigma_k [\Psi]_{kk}. \tag{20}$$

Another relatively easy-to-handle upper bound is

$$\lambda_{\max}\big(\Sigma_{\widetilde{p}}^{1/2}\Psi_{\widetilde{p}}\Sigma_{\widetilde{p}}^{1/2}\big) \leq \|\Sigma_{\widetilde{p}}^{1/2}\Psi_{\widetilde{p}}\Sigma_{\widetilde{p}}^{1/2}\|_\infty = \sup_{k>p} \sum_{r>p} \sqrt{\sigma_k}\sqrt{\sigma_r}\,\big|[\Psi]_{kr}\big|. \tag{21}$$

These bounds can be used, in combination with appropriate block partitioning of $\Sigma_{\widetilde{p}}^{1/2}\Psi_{\widetilde{p}}\Sigma_{\widetilde{p}}^{1/2}$, to provide sharp bounds on the maximal eigenvalue. Block partitioning is useful due to the following: for a positive semidefinite matrix $M = \begin{pmatrix} A_1 & C \\ C^T & A_2 \end{pmatrix}$, we have $\lambda_{\max}(M) \le \lambda_{\max}(A_1) + \lambda_{\max}(A_2)$. We leave the the details on the application of these ideas to examples in Section 3.2.

3.2. Some illustrative examples

Theorem 1 has a number of concrete consequences for different Hilbert spaces and linear operators, and we illustrate a few of them in the following subsections.

3.2.1. Random domain sampling

We begin by stating a corollary of Theorem 1 in application to random time sampling in a reproducing kernel Hilbert space (RKHS). Recall from equation (12) the time sampling operator $\mathbb{S}_{x_1^n}$, and assume that the sample points $\{x_1, \dots, x_n\}$ are drawn in an i.i.d. manner according to some distribution $\mathbb{P}$ on $\mathcal{X}$. Let us further assume that the eigenfunctions ψ_k, $k \ge 1$ are uniformly bounded[7] on $\mathcal{X}$, meaning that

$$\sup_{k\ge1} \sup_{x\in\mathcal{X}} |\psi_k(x)| \le C_\psi. \tag{22}$$

Finally, we assume that $\|\sigma\|_1 := \sum_{k=1}^\infty \sigma_k < \infty$, and that

$$\sigma_{pk} \le C_\sigma\, \sigma_k\, \sigma_p, \quad \text{for some positive constant } C_\sigma \text{ and for all large } p, \tag{23}$$

$$\textstyle\sum_{k>p^m} \sigma_k \le \sigma_p, \quad \text{for some positive integer } m \text{ and for all large } p. \tag{24}$$

Let m_σ be the smallest m for which (24) holds. These conditions on $\{\sigma_k\}$ are satisfied, for example, for both a polynomial decay $\sigma_k = \mathcal{O}(k^{-\alpha})$ with $\alpha > 1$ and an exponential decay $\sigma_k = \mathcal{O}(\rho^k)$ with $\rho \in (0,1)$. In particular, for the polynomial decay, using the tail bound (B.1) in Appendix Appendix B, we can take $m_\sigma = \lceil \frac{\alpha}{\alpha-1} \rceil$ to satisfy (24). For the exponential decay, we can take $m_\sigma = 1$ for $\rho \in (0, \frac{1}{2})$ and $m_\sigma = 2$ for $\rho \in (\frac{1}{2}, 1)$ to satisfy (24).

Define the function

$$\mathcal{G}_n(\varepsilon) := \frac{1}{\sqrt{n}} \sqrt{\sum_{j=1}^\infty \min\{\sigma_j, \varepsilon^2\}}, \tag{25}$$

[7] One can replace $\sup_{x\in\mathcal{X}}$ with essential supremum with respect to $\mathbb{P}$.

as well as the *critical radius*

$$r_n := \inf\{\varepsilon > 0 \,:\, \mathcal{G}_n(\varepsilon) \le \varepsilon^2\}. \tag{26}$$

Corollary 1. *Suppose that $r_n > 0$ and $64\, C_\psi^2\, m_\sigma\, r_n^2 \log(2nr_n^2) \le 1$. Then for any $\varepsilon^2 \in [r_n^2, \sigma_1)$, we have*

$$\mathbb{P}\Big[R_{\mathbb{S}_{x_1^n}}(\varepsilon) > (\widetilde{C}_\psi + \widetilde{C}_\sigma)\,\varepsilon^2\Big] \le 2\exp\Big(-\frac{1}{64\, C_\psi^2\, r_n^2}\Big), \tag{27}$$

where $\widetilde{C}_\psi := 2(1+C_\psi)^2$ and $\widetilde{C}_\sigma := 3(1+C_\psi^{-1})C_\sigma\|\sigma\|_1 + 1$.

We provide the proof of this corollary in Appendix Appendix A. As a concrete example consider a polynomial decay $\sigma_k = \mathcal{O}(k^{-\alpha})$ for $\alpha > 1$, which satisfies assumptions on $\{\sigma_k\}$. Using the tail bound (B.1) in Appendix Appendix B, one can verify that $r_n^2 = \mathcal{O}(n^{-\alpha/(\alpha+1)})$. Note that, in this case,

$$r_n^2 \log(2nr_n^2) = \mathcal{O}(n^{-\frac{\alpha}{\alpha+1}} \log n^{\frac{1}{\alpha+1}}) = \mathcal{O}(n^{-\frac{\alpha}{\alpha+1}} \log n) \to 0, \quad n \to \infty.$$

Hence conditions of Corollary 1 are met for sufficiently large n. It follows that for some constants C_1, C_2 and C_3, we have

$$R_{\mathbb{S}_{x_1^n}}(C_1 n^{-\frac{\alpha}{2(\alpha+1)}}) \le C_2\, n^{-\frac{\alpha}{\alpha+1}}$$

with probability $1 - 2\exp(-C_3 n^{\frac{\alpha}{\alpha+1}})$ for sufficiently large n.

3.2.2. Sobolev kernel

Consider the kernel $\mathbb{K}(x,y) = \min(x,y)$ defined on $\mathcal{X}^2$ where $\mathcal{X} = [0,1]$. The corresponding RKHS is of Sobolev type and can be expressed as

$$\big\{f \in L^2(\mathcal{X}) \;|\; f \text{ is absolutely continuous, } f(0) = 0 \text{ and } f' \in L^2(\mathcal{X})\big\}.$$

Also consider a uniform domain sampling operator $\mathbb{S}_{x_1^n}$, that is, that of (12) with $x_i = i/n, i \le n$ and let $\mathbb{P}$ be uniform (i.e., the Lebesgue measure restricted to $[0,1]$).

This setting has the benefit that many interesting quantities can be computed explicitly, while also having some practical appeal. The following can be shown about the eigen-decomposition of the integral operator $I_{\mathbb{K}}$ introduced in Section 2,

$$\sigma_k = \Big[\frac{(2k-1)\pi}{2}\Big]^{-2}, \quad \psi_k(x) = \sqrt{2}\sin\big(\sigma_k^{-1/2}x\big), \quad k = 1, 2, \dots.$$

In particular, the eigenvalues decay as $\sigma_k = \mathcal{O}(k^{-2})$.

To compute the Ψ, we write

$$[\Psi]_{kr} = \langle \psi_k, \psi_r \rangle_\Phi = \frac{1}{n}\sum_{\ell=1}^{n}\Big\{\cos\frac{(k-r)\ell\pi}{n} - \cos\frac{(k+r-1)\ell\pi}{n}\Big\}. \tag{28}$$

We note that Ψ is periodic in k and r with period $2n$. It is easily verified that $n^{-1}\sum_{\ell=1}^{n}\cos(q\ell\pi/n)$ is equal to -1 for odd values of q and zero for even values, other than $q = 0, \pm 2n, \pm 4n, \dots$. It follows that

$$[\Psi]_{kr} = \begin{cases} 1 + \frac{1}{n} & \text{if } k - r = 0, \\ -1 - \frac{1}{n} & \text{if } k + r = 2n+1\ , \\ \frac{1}{n}(-1)^{k-r} & \text{otherwise} \end{cases} \tag{29}$$

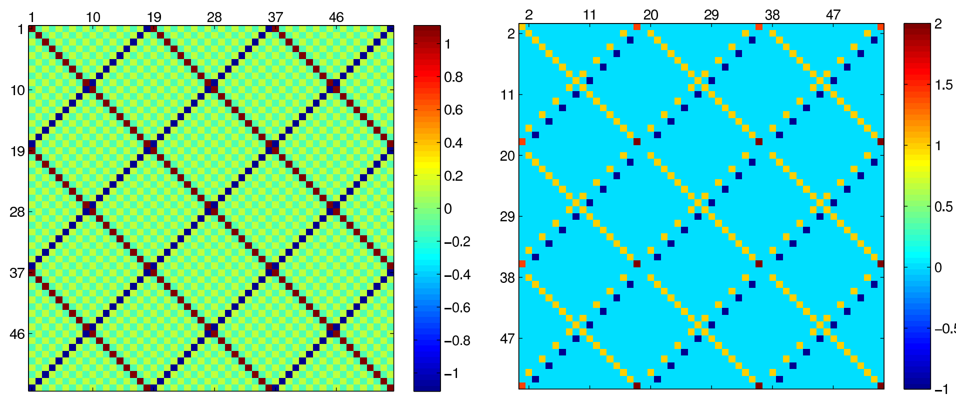
for $1 \le k, r \le 2n$. Letting $\mathbb{I}_s \in \mathbb{R}^n$ be the vector with entries, $(\mathbb{I}_s)_j = (-1)^{j+1}, j \le n$, we observe that $\Psi_n = I_n + \frac{1}{n}\mathbb{I}_s\mathbb{I}_s^T$. It follows that $\lambda_{\min}(\Psi_n) = 1$. It remains to bound the terms in (17) involving the infinite sub-block $\Psi_{\widetilde{n}}$.

The Ψ matrix of this example, given by (29), shares certain properties with the Ψ obtained in other situations involving periodic eigenfunctions $\{\psi_k\}$. We abstract away these properties by introducing a class of periodic Ψ matrices. We call $\Psi_{\widetilde{n}}$ a *sparse periodic* matrix, if each row (or column) is periodic and in each period only a vanishing fraction of elements are large. More precisely, $\Psi_{\widetilde{n}}$ is *sparse periodic* if there exist positive integers γ and η, and positive constants c_1 and c_2, all independent of n, such that each row of $\Psi_{\widetilde{n}}$ is periodic with period γn. and for any row k, there exits a subset of elements $S_k = \{\ell_1, \dots, \ell_\eta\} \subset \{1, \dots, \gamma n\}$ such that

$$\big|[\Psi]_{k,n+r}\big| \le c_1, \qquad r \in S_k, \tag{30a}$$

$$\big|[\Psi]_{k,n+r}\big| \le c_2\, n^{-1}, \qquad r \in \{1, \dots, \gamma n\} \setminus S_k, \tag{30b}$$

The elements of S_k could depend on k, but the cardinality of this set should be the constant η, independent of k and n. Also, note that we are indexing rows and columns of $\Psi_{\widetilde{n}}$ by $\{n{+}1, n{+}2, \dots\}$; in particular, $k \ge n{+}1$. For this class, we have the following whose proof can be found in Appendix Appendix B.



(a) (b)

Figure 2: Sparse periodic Ψ matrices. Display (a) is a plot of the N-by-N leading principal submatrix of Ψ for the Sobolev kernel $(s,t) \mapsto \min\{s,t\}$. Here $n = 9$ and $N = 6n$; the period is $2n = 18$. Display (b) is a the same plot for a Fourier-type kernel. The plots exhibit sparse periodic patterns as defined in Section 3.2.2.

Lemma 1. *Assume $\Psi_{\widetilde{n}}$ to be sparse periodic as defined above and $\sigma_k = \mathcal{O}(k^{-\alpha})$, $\alpha \geq 2$. Then,*

(a) for $\alpha > 2$, $\lambda_{\max}\big(\Sigma_{\widetilde{n}}^{1/2}\Psi_{\widetilde{n}}\Sigma_{\widetilde{n}}^{1/2}\big) = \mathcal{O}(n^{-\alpha})$, $n \to \infty$,

(b) for $\alpha = 2$, $\lambda_{\max}\big(\Sigma_{\widetilde{n}}^{1/2}\Psi_{\widetilde{n}}\Sigma_{\widetilde{n}}^{1/2}\big) = \mathcal{O}(n^{-2}\log n)$, $n \to \infty$.

In particular (29) implies that $\Psi_{\widetilde{n}}$ is sparse periodic with parameters $\gamma = 2$, $\eta = 2$, $c_1 = 2$ and $c_2 = 1$. Hence, part (b) of Lemma 1 applies. Now, we can use (17) with $p = n$ to obtain

$$R_{\mathbb{S}_{x_1^n}}(\varepsilon) \leq 2\varepsilon^2 + \mathcal{O}\big(n^{-2}\log n\big) \tag{31}$$

where we have also used $(a+b)^2 \leq 2a^2 + 2b^2$.

3.2.3. Fourier-type kernels

In this example, we consider an RKHS of functions on $\mathcal{X} = [0,1] \subset \mathbb{R}$, generated by a *Fourier-type* kernel defined as $\mathbb{K}(x,y) := \kappa(x-y)$, $x, y \in [0,1]$, where

$$\kappa(x) = \zeta_0 + \sum_{k=1}^{\infty} 2\zeta_k \cos(2\pi k x), \quad x \in [-1,1]. \tag{32}$$

We assume that (ζ_k) is a $\mathbb{R}_+$-valued nonincreasing sequence in ℓ_1, i.e. $\sum_k \zeta_k < \infty$. Thus, the trigonometric series in (32) is absolutely (and uniformly) convergent. As for the operator Φ, we consider the uniform time sampling operator $\mathbb{S}_{x_1^n}$, as in the previous example. That is, the operator defined in (12) with $x_i = i/n, i \le n$. We take $\mathbb{P}$ to be uniform.

This setting again has the benefit of being simple enough to allow for explicit computations while also practically important. One can argue that the eigen-decomposition of the kernel integral operator is given by

$$\psi_1 = \psi_0^{(c)}, \quad \psi_{2k} = \psi_k^{(c)}, \quad \psi_{2k+1} = \psi_k^{(s)}, \quad k \ge 1 \tag{33}$$

$$\sigma_1 = \zeta_0, \qquad \sigma_{2k} = \zeta_k, \qquad \sigma_{2k+1} = \zeta_k, \qquad k \ge 1 \tag{34}$$

where $\psi_0^{(c)}(x) := 1$, $\psi_k^{(c)}(x) := \sqrt{2}\cos(2\pi k x)$ and $\psi_k^{(s)}(t) := \sqrt{2}\sin(2\pi k x)$ for $k \ge 1$.

For any integer k, let $(\!(k)\!)_n$ denote k modulo n. Also, let $k \mapsto \delta_k$ be the function defined over integers which is 1 at $k = 0$ and zero elsewhere. Let $\iota := \sqrt{-1}$. Using the identity $n^{-1}\sum_{\ell=1}^n \exp(\iota 2\pi k \ell/n) = \delta_{(\!(k)\!)_n}$, one obtains the following,

$$\langle \psi_k^{(c)}, \psi_j^{(c)} \rangle_\Phi = \left[\delta_{(\!(k-j)\!)_n} + \delta_{(\!(k+j)\!)_n}\right]\left(\frac{1}{\sqrt{2}}\right)^{\delta_k + \delta_j}, \tag{35a}$$

$$\langle \psi_k^{(s)}, \psi_j^{(s)} \rangle_\Phi = \delta_{(\!(k-j)\!)_n} - \delta_{(\!(k+j)\!)_n}, \tag{35b}$$

$$\langle \psi_k^{(c)}, \psi_j^{(s)} \rangle_\Phi = 0, \qquad \text{valid for all } j, k \ge 0. \tag{35c}$$

It follows that $\Psi_n = I_n$ if n is odd and $\Psi_n = \text{diag}\{1, 1, \dots, 1, 2\}$ if n is even. In particular, $\lambda_{\min}(\Psi_n) = 1$ for all $n \ge 1$. It is also clear that the principal submatrix of Ψ on indices $\{2, 3, \dots\}$ has periodic rows and columns with period $2n$. If follows that Ψ_n is sparse periodic as defined in Section 3.2.2 with parameters $\gamma = 2$, $\eta = 2$, $c_1 = 2$ and $c_2 = 0$.

Suppose for example that the eigenvalues decay polynomially, say as $\zeta_k = \mathcal{O}(k^{-\alpha})$ for $\alpha > 2$. Then, applying (17) with $p = n$, in combination with Lemma 1 part (a), we get

$$R_{\mathbb{S}_{x_1^n}}(\varepsilon) \le 2\varepsilon^2 + \mathcal{O}(n^{-\alpha}). \tag{36}$$

As another example, consider the exponential decay $\zeta_k = \rho^k$, $k \ge 1$ for some $\rho \in (0, 1)$, which corresponds to the Poisson kernel. In this case, the tail sum of $\{\sigma_k\}$ decays as the sequence itself, namely, $\sum_{k>n} \sigma_k \leq 2 \sum_{k>n} \rho^k = \frac{2\rho}{1-\rho}\rho^k$. Hence, we can simply use the trace bound (20) together with (17) to obtain

$$R_{\mathbb{S}_{x_1^n}}(\varepsilon) \leq 2\varepsilon^2 + \mathcal{O}(\rho^n). \tag{37}$$

4. Proof of Theorem 1

We now turn to the proof of our main theorem. Recall from Section 2.1 the correspondence between any $f \in \mathcal{H}$ and a sequence $\alpha \in \ell_2$; also, recall the diagonal operator $\Sigma : \ell_2 \to \ell_2$ defined by the matrix $\mathrm{diag}\{\sigma_1, \sigma_2, \ldots\}$. Using the definition of (15) of the Ψ matrix, we have

$$\|f\|_\Phi^2 = \langle \alpha, \Sigma^{1/2}\Psi\Sigma^{1/2}\alpha\rangle_{\ell_2},$$

By definition (6) of the Hilbert space $\mathcal{H}$, we have $\|f\|_{\mathcal{H}}^2 = \sum_{k=1}^\infty \alpha_k^2$ and $\|f\|_{L^2}^2 = \sum_k \sigma_k \alpha_k^2$. Letting $B_{\ell_2} = \{\alpha \in \ell_2 \mid \|\alpha\|_{\ell_2} \leq 1\}$ be the unit ball in ℓ_2, we conclude that R_Φ can be written as

$$R_\Phi(\varepsilon) = \sup_{\alpha \in B_{\ell_2}} \big\{Q_2(\alpha) \mid Q_\Phi(\alpha) \leq \varepsilon^2\big\}, \tag{38}$$

where we have defined the quadratic functionals

$$Q_2(\alpha) := \langle \alpha, \Sigma\alpha\rangle_{\ell_2}, \quad \text{and} \quad Q_\Phi(\alpha) := \langle \alpha, \Sigma^{1/2}\Psi\Sigma^{1/2}\alpha\rangle_{\ell_2}. \tag{39}$$

Also let us define the symmetric bilinear form

$$B_\Phi(\alpha, \beta) := \langle \alpha, \Sigma^{1/2}\Psi\Sigma^{1/2}\beta\rangle_{\ell_2}, \quad \alpha, \beta \in \ell^2, \tag{40}$$

whose diagonal is $B_\Phi(\alpha, \alpha) = Q_\Phi(\alpha)$.

We now upper bound $R_\Phi(\varepsilon)$ using a truncation argument. Define the set

$$\mathcal{C} := \{\alpha \in B_{\ell_2} \mid Q_\Phi(\alpha) \leq \varepsilon^2\}, \tag{41}$$

corresponding to the feasible set for the optimization problem (38). For each integer $p = 1, 2, \ldots$, consider the following truncated sequence spaces

$$\mathcal{T}_p := \big\{\alpha \in \ell_2 \mid \alpha_i = 0, \quad \text{for all } i > p\big\}, \quad \text{and}$$
$$\mathcal{T}_p^\perp := \big\{\alpha \in \ell_2 \mid \alpha_i = 0, \quad \text{for all } i = 1, 2, \ldots p\big\}.$$

Note that ℓ_2 is the direct sum of $\mathcal{T}_p$ and $\mathcal{T}_p^\perp$. Consequently, any fixed $\alpha \in \mathcal{C}$ can be decomposed as $\alpha = \xi + \gamma$ for some (unique) $\xi \in \mathcal{T}_p$ and $\gamma \in \mathcal{T}_p^\perp$. Since Σ is a diagonal operator, we have

$$Q_2(\alpha) = Q_2(\xi) + Q_2(\gamma).$$

Moreover, since any $\alpha \in \mathcal{C}$ is feasible for the optimization problem (38), we have

$$Q_\Phi(\alpha) \ = \ Q_\Phi(\xi) + 2B_\Phi(\xi,\gamma) + Q_\Phi(\gamma) \ \leq \ \varepsilon^2. \tag{42}$$

Note that since $\gamma \in \mathcal{T}_p^\perp$, it can be written as $\gamma = (0_p, c)$, where 0_p is a vector of p zeroes, and $c = (c_1, c_2, \dots) \in \ell_2$. Similarly, we can write $\xi = (x, 0)$ where $x \in \mathbb{R}^p$. Then, each of the terms $Q_\Phi(\xi)$, $B_\Phi(\xi,\gamma)$, $Q_\Phi(\gamma)$ can be expressed in terms of block partitions of $\Sigma^{1/2}\Psi\Sigma^{1/2}$. For example,

$$Q_\Phi(\xi) = \langle x, Ax\rangle_{\mathbb{R}^p}, \quad Q_\Phi(\gamma) = \langle y, Dy\rangle_{\ell_2}, \tag{43}$$

where $A := \Sigma_p^{1/2}\Psi_p\Sigma_p^{1/2}$ and $D := \Sigma_{\widetilde{p}}^{1/2}\Psi_{\widetilde{p}}\Sigma_{\widetilde{p}}^{1/2}$, in correspondence with the block partitioning notation of Appendix Appendix F. We now apply inequality (F.2) derived in Appendix Appendix F. Fix some $\rho^2 \in (0,1)$ and take

$$\kappa^2 := \rho^2 \lambda_{\max}(\Sigma_{\widetilde{p}}^{1/2}\Psi_{\widetilde{p}}\Sigma_{\widetilde{p}}^{1/2}), \tag{44}$$

so that condition (F.5) is satisfied. Then, (F.2) implies

$$Q_\Phi(\xi) + 2B_\Phi(\xi,\gamma) + Q_\Phi(\gamma) \geq \rho^2 Q_\Phi(\xi) - \frac{\kappa^2}{1-\rho^2}\|\gamma\|_2^2. \tag{45}$$

Combining (42) and (45), we obtain

$$Q_\Phi(\xi) \leq \frac{\varepsilon^2}{\rho^2} + \frac{\lambda_{\max}(\Sigma_{\widetilde{p}}^{1/2}\Psi_{\widetilde{p}}\Sigma_{\widetilde{p}}^{1/2})}{1-\rho^2}\|\gamma\|_2^2. \tag{46}$$

We further note that $\|\gamma\|_2^2 \leq \|\gamma\|_2^2 + \|\xi\|_2^2 = \|\alpha\|_2^2 \leq 1$. It follows that

$$Q_\Phi(\xi) \ \leq \ \widetilde{\varepsilon}^2, \quad \text{where} \quad \widetilde{\varepsilon}^2 := \frac{\varepsilon^2}{\rho^2} + \frac{\lambda_{\max}(\Sigma_{\widetilde{p}}^{1/2}\Psi_{\widetilde{p}}\Sigma_{\widetilde{p}}^{1/2})}{1-\rho^2}. \tag{47}$$

Let us define

$$\widetilde{\mathcal{C}} := \{\xi \in B_{\ell_2} \cap \mathcal{T}_p \ | \ Q_\Phi(\xi) \le \widetilde{\varepsilon}^2\}. \tag{48}$$

Then, our arguments so far show that for $\alpha \in \mathcal{C}$,

$$Q_2(\alpha) = Q_2(\xi) + Q_2(\gamma) \ \le \ \underbrace{\sup_{\xi \in \widetilde{\mathcal{C}}} Q_2(\xi)}_{S_p} \ + \ \underbrace{\sup_{\gamma \in B_{\ell_2} \cap \mathcal{T}_p^\perp} Q_2(\gamma)}_{S_p^\perp}. \tag{49}$$

Taking the supremum over $\alpha \in \mathcal{C}$ yields the upper bound

$$R_\Phi(\varepsilon) \le S_p + S_p^\perp.$$

It remains to bound each of the two terms on the right-hand side. Beginning with the term $S_p^\perp$ and recalling the decomposition $\gamma = (0_p, c)$, we have $Q_2(\gamma) = \sum_{k=1}^\infty \sigma_{k+p} c_k^2$, from which it follows that

$$S_p^\perp = \sup \Big\{ \sum_{k=1}^\infty \sigma_{k+p}\, c_k^2 \ \mid \ \sum_{k=1}^\infty c_k^2 \le 1 \Big\} \ = \ \sigma_{p+1},$$

since $\{\sigma_k\}_{k=1}^\infty$ is a nonincreasing sequence by assumption.

We now control the term S_p. Recalling the decomposition $\xi = (x, 0)$ where $x \in \mathbb{R}^p$, we have

$$\begin{aligned}
S_p = \sup_{\xi \in \widetilde{\mathcal{C}}} Q_2(\xi) &= \sup \big\{ \langle x, \Sigma_p\, x\rangle \ : \ \langle x, x\rangle \le 1, \ \langle x, \Sigma_p^{1/2} \Psi_p \Sigma_p^{1/2}\, x\rangle \le \widetilde{\varepsilon}^2 \big\} \\
&= \sup_{\langle x,x\rangle \le 1} \ \inf_{t \ge 0} \big\{ \langle x, \Sigma_p x\rangle + t\big(\widetilde{\varepsilon}^2 - \langle x, \Sigma_p^{1/2}\Psi_p \Sigma_p^{1/2}\, x\rangle\big) \big\} \\
&\overset{(a)}{\le} \inf_{t \ge 0} \big\{ \sup_{\langle x,x\rangle \le 1} \langle x, \Sigma_p^{1/2}(I_p - t\Psi_p)\Sigma_p^{1/2}\, x\rangle + t\, \widetilde{\varepsilon}^2 \big\}
\end{aligned}$$

where inequality (a) follows by Lagrange (weak) duality. It is not hard to see that for any symmetric matrix M, one has

$$\sup \big\{ \langle x, Mx\rangle \ : \ \langle x, x\rangle \le 1 \big\} = \max \big\{ 0, \lambda_{\max}(M) \big\}.$$

Putting the pieces together and optimizing over ρ^2, noting that

$$\inf_{r \in (0,1)} \Big\{ \frac{a}{r} + \frac{b}{1-r} \Big\} = (\sqrt{a} + \sqrt{b})^2$$

for any $a, b > 0$, completes the proof of the bound (16).

We now prove bound (17), using the same decomposition and notation established above, but writing an upper bound on $Q_2(\alpha)$slightly different form (49). In particular, the argument leading to (49), also shows that

$$R_\Phi(\varepsilon) \le \sup_{\xi \in \mathcal{T}_p,\, \gamma \in \mathcal{T}_p^\perp} \big\{ Q_2(\xi) + Q_2(\gamma) \;\big|\; \xi + \gamma \in B_{\ell_2},\; Q_\Phi(\xi) \le \widetilde{\varepsilon}^2 \big\}. \tag{50}$$

Recalling the expression (39) for $Q_\Phi(\xi)$ and noting that $\Psi_p \succeq \lambda_{\min}(\Psi_p) I_p$ implies $A = \Sigma_p^{1/2} \Psi_p \Sigma_p^{1/2} \succeq \lambda_{\min}(\Psi_p)\Sigma_p$, we have

$$Q_\Phi(\xi) \;\ge\; \lambda_{\min}(\Psi_p)\, Q_2(\xi). \tag{51}$$

Now, since we are assuming $\lambda_{\min}(\Psi_p) > 0$, we have

$$R_\Phi(\varepsilon) \le \sup_{\xi \in \mathcal{T}_p,\, \gamma \in \mathcal{T}_p^\perp} \Big\{ Q_2(\xi) + Q_2(\gamma) \;\Big|\; \xi + \gamma \in B_{\ell_2},\; Q_2(\xi) \le \frac{\widetilde{\varepsilon}^2}{\lambda_{\min}(\Psi_p)} \Big\}. \tag{52}$$

The RHS of the above is an instance of the Fourier truncation problem with ε^2 replaced with $\widetilde{\varepsilon}^2/\lambda_{\min}(\Psi_p)$. That problem is workout in detail in Appendix Appendix E. In particular, applying equation (E.1) in Appendix Appendix E with ε^2 changed to $\widetilde{\varepsilon}^2/\lambda_{\min}(\Psi_p)$ completes the proof of (17). Figure 3 provides a graphical representation of the geometry of the proof.

5. Conclusion

We considered the problem of bounding (squared) L^2 norm of functions in a Hilbert unit ball, based on restrictions on an operator-induced norm acting as a surrogate for the L^2 norm. In particular, given that $f \in B_\mathcal{H}$ and $\|f\|_\Phi^2 \le \varepsilon^2$, our results enable us to obtain, by estimating norms of certain finite and infinite dimensional matrices, inequalities of the form

$$\|f\|_{L^2}^2 \le c_1 \varepsilon^2 + h_{\Phi,\mathcal{H}}(\sigma_n)$$

where $\{\sigma_n\}$ are the eigenvalues of the operator embedding $\mathcal{H}$ in L^2, $h_{\Phi,\mathcal{H}}(\cdot)$ is an increasing function (depending on Φ and $\mathcal{H}$) and $c_1 \ge 1$ is some constant. We considered examples of operators Φ (uniform time sampling and Fourier truncation) and Hilbert spaces $\mathcal{H}$ (Sobolev, Fourier-type RKHSs) and showed

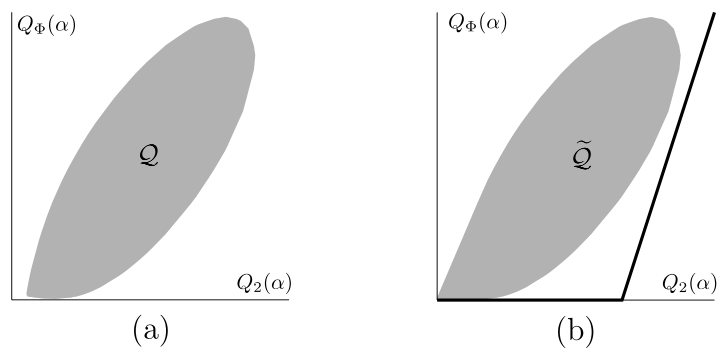


Figure 3: Geometry of the proof of (17). Display (a) is a plot of the set $\mathcal{Q} := \{(Q_2(\alpha), Q_\Phi(\alpha)) : \|\alpha\|_{\ell_2} = 1\} \subset \mathbb{R}^2$. This is a convex set as a consequence of Hausdorff-Toeplitz theorem on convexity of the numerical range and preservation of convexity under projections. Display (b) shows the set $\widetilde{\mathcal{Q}} := \text{conv}(0, \mathcal{Q})$, i.e., the convex hull of $\{0\} \cup \mathcal{Q}$. Observe that $R_\Phi(\varepsilon) = \sup\{x : (x, y) \in \widetilde{\mathcal{Q}},\ y \le \varepsilon^2\}$. For any fixed $r \in (0,1)$, the bound of (17) is a piecewise linear approximation to one side of $\widetilde{\mathcal{Q}}$ as shown in Display (b).

that it is possible to obtain optimal scaling $h_{\Phi,\mathcal{H}}(\sigma_n) = \mathcal{O}(\sigma_n)$ in most of those cases. We also considered random time sampling, under polynomial eigendecay $\sigma_n = \mathcal{O}(n^{-\alpha})$, and effectively showed that $h_{\Phi,\mathcal{H}}(\sigma_n) = \mathcal{O}(n^{-\alpha/(\alpha+1)})$ (for ε small enough), with high probability as $n \to \infty$. This last result complements those on related quantities obtained by techniques form empirical process theory, and we conjecture it to be sharp.

Acknowledgements

AA and MJW were partially supported by NSF Grant CAREER-CCF-0545862 and AFOSR Grant 09NL184.

Appendix A. Analysis of random time sampling

This section is devoted to the proof of Corollary 1 on random time sampling in reproducing kernel Hilbert spaces. The proof is based on an auxiliary result, which we begin by stating. Fix some positive integer m and define

$$\nu(\varepsilon) = \nu(\varepsilon; m) := \inf\Big\{p : \sum_{k > p^m} \sigma_k \le \varepsilon^2\Big\}. \tag{A.1}$$

With this notation, we have

Lemma 2. *Assume* $\varepsilon^2 < \sigma_1$ *and* $32\, C_\psi^2\, m\, \nu(\varepsilon) \log \nu(\varepsilon) \le n$*. Then,*

$$\mathbb{P}\big\{ R_{\mathbb{S}_{x_1^n}}(\varepsilon) > \widetilde{C}_\psi\, \varepsilon^2 + \widetilde{C}_\sigma\, \sigma_{\nu(\varepsilon)} \big\} \le 2 \exp\Big(- \frac{1}{32 C_\psi^2} \frac{n}{\nu(\varepsilon)} \Big). \tag{A.2}$$

We prove this claim in Section Appendix A.2 below.

Appendix A.1. Proof of Corollary 1

To apply the lemma, recall that we assume that there exists m such that for all (large) p, one has

$$\sum_{k > p^m} \sigma_k \le \sigma_p. \tag{A.3}$$

and we let m_σ be the smallest such m. We define

$$\mu(\varepsilon) := \inf \big\{ p : \sigma_p \le \varepsilon^2 \big\}, \tag{A.4}$$

and note that by (A.3), we have $\nu(\varepsilon; m_\sigma) \le \mu(\varepsilon)$. Then, Lemma 2 states that as long as $\varepsilon^2 < \sigma_1$ and $32 C_\psi^2 m_\sigma \mu(\varepsilon) \log \mu(\varepsilon) \le n$, we have

$$\mathbb{P}\big\{ R_{\mathbb{S}_{x_1^n}}(\varepsilon) > (\widetilde{C}_\psi + \widetilde{C}_\sigma) \varepsilon^2 \big\} \le 2 \exp\Big(- \frac{1}{32 C_\psi^2} \frac{n}{\mu(\varepsilon)} \Big). \tag{A.5}$$

Now by the definition of $\mu(\varepsilon)$, we have $\sigma_j > \varepsilon^2$ for $j < \mu(\varepsilon)$, and hence

$$\mathcal{G}_n^2(\varepsilon) \ge \frac{1}{n} \sum_{j \,<\, \mu(\varepsilon)} \min\{\sigma_j, \varepsilon^2\} = \frac{\mu(\varepsilon) - 1}{n}\, \varepsilon^2 \ge \frac{\mu(\varepsilon)}{2n}\, \varepsilon^2,$$

since $\mu(\varepsilon) \ge 2$ when $\varepsilon^2 < \sigma_1$. One can argue that $\varepsilon \mapsto \mathcal{G}_n(\varepsilon)/\varepsilon$ is nonincreasing. It follows from definition (26) that for $\varepsilon \ge r_n$, we have

$$\mu(\varepsilon) \le 2n \Big(\frac{\mathcal{G}(\varepsilon)}{\varepsilon} \Big)^2 \le 2n \Big(\frac{\mathcal{G}(r_n)}{r_n} \Big)^2 \le 2n r_n^2,$$

which completes the proof of Corollary 1.

Appendix A.2. Proof of Lemma 2

For $\xi \in \mathbb{R}^p$, let $\xi \otimes \xi$ be the rank-one operator on $\mathbb{R}^p$ given by $\eta \mapsto \langle \xi, , \eta \rangle_2 \, \xi$. For an operator A on $\mathbb{R}^p$, let $|\!|\!| A |\!|\!|_2$ denote its usual operator norm, $|\!|\!| A |\!|\!|_2 := \sup_{\|x\|_2 \le 1} \|Ax\|_2$. Recall that for a symmetric (i.e., real self-adjoint) operator A on $\mathbb{R}^p$, $|\!|\!| A |\!|\!|_2 = \sup\{|\lambda| : \lambda \text{ an eigenvalue of } A\}$. It follows that $|\!|\!| A |\!|\!|_2 \le \alpha$ is equivalent to $-\alpha I_p \preceq A \preceq \alpha I_p$.

Our approach is to first show that $|\!|\!| \Psi_p - I_p |\!|\!|_2 \le \frac{1}{2}$ for some properly chosen p with high probability. It then follows that $\lambda_{\min}(\Psi_p) \ge \frac{1}{2}$ and we can use bound (17) for that value of p. Then, we need to control $\lambda_{\max}\big(\Sigma_{\widetilde{p}}^{1/2} \Psi_{\widetilde{p}} \Sigma_{\widetilde{p}}^{1/2}\big)$. To do this, we further partition $\Psi_{\widetilde{p}}$ into blocks. In order to have a consistent notation, we look at the whole matrix Ψ and let $\Psi^{(k)}$ be the principal submatrix indexed by $\{(k-1)p+1, \dots, (k-1)p+p\}$, for $k = 1, 2, \dots, p^{m-1}$. Throughout the proof, m is assumed to be a fixed positive integer. Also, let $\Psi^{(\infty)}$ be the principal submatrix of Ψ indexed by $\{p^m+1, p^m+2, \dots\}$. This provides a full partitioning of Ψ for which $\Psi^{(1)}, \dots, \Psi^{(p^{m-1})}$ and $\Psi^{(\infty)}$ are the diagonal blocks, the first p^{m-1} of which are p-by-p matrices and the last an infinite matrix. To connect with our previous notations, we note that $\Psi^{(1)} = \Psi_p$ and that $\Psi^{(2)}, \dots, \Psi^{(p^{m-1})}, \Psi^{(\infty)}$ are diagonal blocks of $\Psi_{\widetilde{p}}$. Let us also partition the Σ matrix and name its diagonal blocks similarly.

We will argue that, in fact, we have $|\!|\!| \Psi^{(k)} - I_p |\!|\!|_2 \le \frac{1}{2}$ for all $k = 1, \dots, p^{m-1}$, with high probability. Let $\mathcal{A}_p$ denote the event on which this claim holds. In particular, on event $\mathcal{A}_p$, we have $\Psi^{(k)} \preceq \frac{3}{2} I_p$ for $k = 2, \dots, p^{m-1}$; hence, we can write

$$\begin{aligned}
\lambda_{\max}\big(\Sigma_{\widetilde{p}}^{1/2} \Psi_{\widetilde{p}} \Sigma_{\widetilde{p}}^{1/2}\big) &\le \sum_{k=2}^{p^{m-1}} \lambda_{\max}\Big(\sqrt{\Sigma^{(k)}}\,\Psi^{(k)}\sqrt{\Sigma^{(k)}}\Big) + \lambda_{\max}\Big(\sqrt{\Sigma^{(\infty)}}\,\Psi^{(\infty)}\sqrt{\Sigma^{(\infty)}}\Big) \\
&\le \frac{3}{2} \sum_{k=2}^{p^{m-1}} \lambda_{\max}\big(\Sigma^{(k)}\big) + \operatorname{tr}\Big(\sqrt{\Sigma^{(\infty)}}\,\Psi^{(\infty)}\sqrt{\Sigma^{(\infty)}}\Big) \\
&= \frac{3}{2} \sum_{k=2}^{p^{m-1}} \sigma_{(k-1)p+1} + \sum_{k > p^m} \sigma_k [\Psi]_{kk}.
\end{aligned} \tag{A.6}$$

Using assumptions (23) on the sequence $\{\sigma_k\}$, the first sum can be bounded as

$$\sum_{k=2}^{p^{m-1}} \sigma_{(k-1)p+1} \le \sum_{k=2}^{p^{m-1}} \sigma_{(k-1)p} \le \sum_{k=2}^{p^{m-1}} C_\sigma \sigma_{k-1} \sigma_p \le C_\sigma \|\sigma\|_1 \sigma_p$$

Using the uniform boundedness assumption (A.1), we have $[\Psi]_{kk} = n^{-1}\sum_{i=1}^n \psi_k^2(x_i) \leq C_\psi^2$. Hence the second sum in (A.6) is bounded above by $C_\psi^2 \sum_{k>p^m} \sigma_k$.

We can now apply Theorem 1. Assume for the moment that $\varepsilon^2 \geq \sum_{k>p^m} \sigma_k$ so that the right-hand side of (A.6) is bounded above by $\frac{3}{2} C_\sigma \|\sigma\|_1 \sigma_p + C_\psi^2 \varepsilon^2$. Applying bound (17), on event $\mathcal{A}_p$, with[8] $r = (1+C_\psi)^{-1}$, we get

$$
\begin{aligned}
R_{\mathbb{S}_{x_1^n}}(\varepsilon^2) &\leq 2\Big\{ r^{-1}\varepsilon^2 + (1-r)^{-1}\Big(\frac{3}{2} C_\sigma \|\sigma\|_1 \sigma_p + C_\psi^2 \varepsilon^2\Big)\Big\} + \sigma_{p+1} \\
&= 2(1+C_\psi)^2 \varepsilon^2 + 3(1+C_\psi^{-1}) C_\sigma \|\sigma\|_1 \sigma_p + \sigma_{p+1}. \\
&\leq \widetilde{C}_\psi\, \varepsilon^2 + \widetilde{C}_\sigma\, \sigma_p
\end{aligned}
$$

where $\widetilde{C}_\psi := 2(1+C_\psi)^2$ and $\widetilde{C}_\sigma := 3(1+C_\psi^{-1}) C_\sigma \|\sigma\|_1 + 1$. To summarize, we have shown the following

$$
\text{Event } \mathcal{A}_p \quad \text{and} \quad \varepsilon^2 \geq \sum_{k>p^m} \sigma_k \implies R_{\mathbb{S}_{x_1^n}}(\varepsilon^2) \leq \widetilde{C}_\psi\, \varepsilon^2 + \widetilde{C}_\sigma\, \sigma_p. \tag{A.7}
$$

It remains to control the probability of $\mathcal{A}_p := \bigcap_{k=1}^{p^m-1} \big\{ \|\!| \Psi^{(k)} - I_p |\!\|_2 \leq \frac{1}{2} \big\}$. We start with the deviation bound on $\Psi^{(1)} - I_p$, and then extend by union bound. We will use the following lemma which follows, for example, from the Ahlswede-Winter bound [8], or from [9]. (See also [10, 11, 12].)

Lemma 3. *Let $\xi_1, \ldots, \xi_n$ be i.i.d. random vectors in $\mathbb{R}^p$ with $\mathbb{E}\, \xi_1 \otimes \xi_1 = I_p$ and $\|\xi_1\|_2 \leq C_p$ almost surely for some constant C_p. Then, for $\delta \in (0,1)$,*

$$
\mathbb{P}\Big\{ \Big|\!\Big|\!\Big| n^{-1} \sum_{i=1}^n \xi_i \otimes \xi_i - I_p \Big|\!\Big|\!\Big|_2 > \delta \Big\} \leq p \exp\big(-\frac{n\delta^2}{4C_p^2} \big). \tag{A.8}
$$

Recall that for the time sampling operator, $[\Phi\, \psi_k]_i = \frac{1}{\sqrt{n}} \psi_k(x_i)$ so that from (15),

$$
\Psi_{k\ell} = \frac{1}{n} \sum_{i=1}^n \psi_k(x_i) \psi_\ell(x_i)
$$

[8] We are using the alternate form of the bound based on $(\sqrt{A} + \sqrt{B})^2 = \inf_{r \in (0,1)} \big\{ A r^{-1} + B(1-r)^{-1} \big\}$.

Let $\xi_i := (\psi_k(x_i), 1 \le k \le p) \in \mathbb{R}^p$ for $i = 1, \dots, n$. Then, $\{\xi_i\}$ satisfy the conditions of Lemma 3. In particular, letting e_k denote the k-th standard basis vector of $\mathbb{R}^p$, we note that

$$\langle e_k, \mathbb{E}(\xi_i \otimes \xi_i) e_\ell \rangle_2 = \mathbb{E}\langle e_k, \xi_i \rangle_2 \langle e_\ell, \xi_i \rangle_2 = \langle \psi_k, \psi_\ell \rangle_{L^2} = \delta_{k\ell}$$

and $\|\xi_i\|_2 \le \sqrt{p}\, C_\psi$, where we have used uniform boundedness of $\{\psi_k\}$ as in (22). Furthermore, we have $\Psi^{(1)} = n^{-1} \sum_{i=1}^n \xi_i \otimes \xi_i$. Applying Lemma 3 with $C_p = \sqrt{p} C_\psi$ yields,

$$\mathbb{P}\big\{ \| \Psi^{(1)} - I_p \|_2 > \delta \big\} \;\le\; p \exp\big(-\frac{\delta^2}{4C_\psi^2} \frac{n}{p} \big). \tag{A.9}$$

Similar bounds hold for $\Psi^{(k)}$, $k = 2, \dots, p^{m-1}$. Applying the union bound, we get

$$\mathbb{P} \bigcup_{k=1}^{p^{m-1}} \big\{ \| \Psi^{(k)} - I_p \|_2 > \delta \big\} \;\le\; \exp\big(m \log p - \frac{\delta^2}{4C_\psi^2} \frac{n}{p} \big).$$

For simplicity, let $A = A_{n,p} := n/(4C_\psi^2\, p)$. We impose $m \log p \le \frac{A}{2} \delta^2$ so that the exponent in (A.9) is bounded above by $-\frac{A}{2}\delta^2$. Furthermore, for our purpose, it is enough to take $\delta = \frac{1}{2}$. It follows that

$$\mathbb{P}(\mathcal{A}_p^c) = \mathbb{P} \bigcup_{k=1}^{p^{m-1}} \big\{ \| \Psi^{(k)} - I_p \|_2 > \frac{1}{2} \big\} \;\le\; \exp\big(-\frac{1}{32 C_\psi^2} \frac{n}{p} \big), \tag{A.10}$$

if $32 C_\psi^2\, m\, p \log p \le n$. Now, by (A.7), under $\varepsilon^2 \ge \sum_{k > p^m} \sigma_k$, $R_{\mathbb{S}_{x_1^n}}(\varepsilon^2) > \widetilde{C}_\psi\, \varepsilon^2 + \widetilde{C}_\sigma\, \sigma_p$ implies $\mathcal{A}_p^c$. Thus, the exponential bound in (A.10) holds for $\mathbb{P}\{R_{\mathbb{S}_{x_1^n}}(\varepsilon^2) > \widetilde{C}_\psi\, \varepsilon^2 + \widetilde{C}_\sigma\, \sigma_p\}$ under the assumptions. We are to choose p and the bound is optimized by making p as small as possible. Hence, we take p to be $\nu(\varepsilon) := \inf\{p : \ \varepsilon^2 \ge \sum_{k > p^m} \sigma_k\}$ which proves Lemma 2. (Note that, in general, $\nu(\varepsilon)$ takes its values in $\{0, 1, 2, \dots\}$. The assumption $\varepsilon^2 < \sigma_1$ guarantees that $\nu(\varepsilon) \ne 0$.)

Appendix B. Proof of Lemma 1

Assume $\sigma_k = C k^{-\alpha}$, for some $\alpha \ge 2$. First, note the following upper bound on the tail sum

$$\sum_{k > p} \sigma_k \le C \int_p^\infty x^{-\alpha}\, dx = C_1(\alpha)\, p^{1-\alpha}. \tag{B.1}$$

Furthermore, from the bounds (30a) and (30b), we have, for $k \geq n+1$,

$$[\Psi]_{kk} \leq \min\{c_1, c_2\}. \tag{B.2}$$

To simplify notation, let us define $I_n := \{1, 2, \ldots, \gamma n\}$.

Consider the case $\alpha > 2$. We will use the ℓ_∞–ℓ_∞ upper bound of (21), with $p = n$. Fix some $k \geq n+1$. Note that $\sigma_k \leq \sigma_{n+1}$. Then, recalling the assumptions on Ψ and the definition of S_k, we have

$$\begin{aligned}
\sum_{\ell \geq n+1} \sqrt{\sigma_k}\sqrt{\sigma_\ell}\,\big|[\Psi]_{k,\ell}\big| &\leq \sqrt{\sigma_{n+1}} \sum_{q=0}^{\infty} \sum_{r=1}^{\gamma n} \sqrt{\sigma_{n+r+q\gamma n}}\big|[\Psi]_{k,n+r+q\gamma n}\big| \\
&= \sqrt{\sigma_{n+1}} \sum_{q=0}^{\infty} \sum_{r=1}^{\gamma n} \sqrt{\sigma_{n+r+q\gamma n}}\big|[\Psi]_{k,n+r}\big| \\
&\leq \sqrt{\sigma_{n+1}} \sum_{q=0}^{\infty} \Big\{ c_1 \sum_{r \in S_k} \sqrt{\sigma_{n+r+q\gamma n}} + \frac{c_2}{n} \sum_{r \in I_n \backslash S_k} \sqrt{\sigma_{n+r+q\gamma n}} \Big\}.
\end{aligned} \tag{B.3}$$

Using (B.1), the second double sum in (B.3) is bounded by

$$\sum_{q=0}^{\infty} \sum_{r \in I_n \backslash S_k} \sqrt{\sigma_{n+r+q\gamma n}} \;\leq\; \sum_{\ell > n} \sqrt{\sigma_\ell} \;\leq\; C_2(\alpha)\, n^{1-\alpha/2}. \tag{B.4}$$

Recalling that $S_k \subset I_n$ and $|S_k| = \eta$, the first double sum in (B.3) can be bounded as follows

$$\begin{aligned}
\sum_{q=0}^{\infty} \sum_{r \in S_k} \sqrt{\sigma_{n+r+q\gamma n}} &= \sqrt{C} \sum_{q=0}^{\infty} \sum_{r \in S_k} (n + r + q\gamma n)^{-\alpha/2} \\
&\leq \sqrt{C} \sum_{q=0}^{\infty} \sum_{r \in S_k} (n + q\gamma n)^{-\alpha/2} \\
&\leq \sqrt{C}\, \eta \sum_{q=0}^{\infty} (1 + q\gamma)^{-\alpha/2} n^{-\alpha/2} \\
&\leq \sqrt{C}\, \eta \Big(1 + \gamma^{-\alpha/2} \sum_{q=1}^{\infty} q^{-\alpha/2}\Big) n^{-\alpha/2} \\
&= C_3(\alpha, \gamma, \eta)\, n^{-\alpha/2}
\end{aligned} \tag{B.5}$$

where in the last line we have used $\sum_{q=1}^{\infty} q^{-\alpha/2} < \infty$ due to $\alpha/2 > 1$. Combining (B.3), (B.4) and (B.5) and noting that $\sqrt{\sigma_{n+1}} \leq \sqrt{C} n^{-\alpha/2}$, we obtain

$$\sum_{\ell \geq n+1} \sqrt{\sigma_k}\sqrt{\sigma_\ell}\,\big|[\Psi]_{k,\ell}\big| \leq \sqrt{C} n^{-\alpha/2}\Big\{c_1 C_3(\alpha,\gamma,\eta)\, n^{-\alpha/2} + \frac{c_2}{n}\, C_2(\alpha)\, n^{1-\alpha/2}\Big\} = C_4(\alpha,\eta,\gamma)\, n^{-\alpha}. \tag{B.6}$$

Taking supremum over $k \geq 1$ and applying the ℓ_∞–ℓ_∞ bound of (21), with $p = n$, concludes the proof of part (a).

Now, consider the case $\alpha = 2$. The above argument breaks down in this case because $\sum_{q=1}^{\infty} q^{-\alpha/2}$ does not converge for $\alpha = 2$. A remedy is to further partition the matrix $\Sigma_{\widetilde{n}}^{1/2} \Psi_{\widetilde{n}} \Sigma_{\widetilde{n}}^{1/2}$. Recall that the rows and columns of this matrix are indexed by $\{n+1, n+2, \dots\}$. Let A be the principal submatrix indexed by $\{n{+}1, n{+}2, \dots, n^2\}$ and D be the principal submatrix indexed by $\{n^2{+}1, n^2{+}2, \dots\}$. We will use a combination of the bounds (30a) and (30b), and the well-known perturbation bound $\lambda_{\max}\big[\big(\begin{smallmatrix} A & C \\ C^T & D \end{smallmatrix}\big)\big] \leq \lambda_{\max}(A) + \lambda_{\max}(D)$, to write

$$\lambda_{\max}\big(\Sigma_{\widetilde{n}}^{1/2} \Psi_{\widetilde{n}} \Sigma_{\widetilde{n}}^{1/2}\big) \leq \lambda_{\max}(A) + \lambda_{\max}(D) \leq |||A|||_\infty + \mathrm{tr}(D). \tag{B.7}$$

The second term is bounded as

$$\mathrm{tr}(D) = \sum_{k > n^2} \sigma_k\, [\Psi]_{kk} \leq \min\{c_1, c_2\} \sum_{k > n^2} \sigma_k = \min\{c_1, c_2\}\, (n^2)^{1-2} = C_5(\gamma)\, n^{-2}, \tag{B.8}$$

where we have used (B.1) and (B.2). To bound the first term, fix $k \in \{n+1, \dots, n^2\}$. By an argument similar to that of part (a) and noting that $\gamma \geq 1$, hence $\gamma n^2 \geq n^2$, we have

$$\begin{aligned} \sum_{\ell = n+1}^{n^2} \sqrt{\sigma_k}\sqrt{\sigma_\ell}\,\big|[\Psi]_{k,\ell}\big| &\leq \sqrt{\sigma_{n+1}} \sum_{q=0}^{n} \sum_{r=1}^{\gamma n} \sqrt{\sigma_{n+r+q\gamma n}}\big|[\Psi]_{k,n+r}\big| \\ &\leq \sqrt{\sigma_{n+1}} \sum_{q=0}^{n} \Big\{ c_1 \sum_{r \in S_k} \sqrt{\sigma_{n+r+q\gamma n}} + \frac{c_2}{n} \sum_{r \in I_n \setminus S_k} \sqrt{\sigma_{n+r+q\gamma n}} \Big\}. \end{aligned} \tag{B.9}$$

Using $\gamma \geq 1$ again, the second double sum in (B.9) is bounded as

$$\sum_{q=0}^{n} \sum_{r \in I_n \backslash S_k} \sqrt{\sigma_{n+r+q\gamma n}} \leq \sum_{\ell=n+1}^{3\gamma n^2} \sqrt{\sigma_\ell} \leq \sqrt{C} \sum_{\ell=2}^{3\gamma n^2} \frac{1}{\ell} \leq \sqrt{C} \log(3\gamma n^2) \leq C_6(\gamma) \log n, \tag{B.10}$$

for sufficiently large n. Note that we have used the bound $\sum_{\ell=2}^{p} \ell^{-1} \leq \int_1^p x^{-1}\, dx = \log p$. The first double sum in (B.9) is bounded as follows

$$\begin{aligned}
\sum_{q=0}^{\infty} \sum_{r \in S_k} \sqrt{\sigma_{n+r+q\gamma n}} &= \sqrt{C} \sum_{q=0}^{n} \sum_{r \in S_k} (n+r+q\gamma n)^{-1} \\
&\leq \sqrt{C}\, \eta \sum_{q=0}^{n} (1+q\gamma)^{-1} n^{-1} \\
&\leq \sqrt{C}\, \eta \Big(1 + \gamma^{-1} + \gamma^{-1} \sum_{q=2}^{n} q^{-1}\Big) n^{-1} \\
&= C_7(\gamma, \eta)\, n^{-1} \log n,
\end{aligned} \tag{B.11}$$

for n sufficiently large. Combining (B.9), (B.10) and (B.11), taking supremum over k and using the simple bound $\sqrt{\sigma_{n+1}} \leq \sqrt{C} n^{-1}$, we get

$$|||A|||_\infty \leq \sqrt{C} n^{-1} \Big\{ c_1 C_7(\gamma, \eta) \frac{\log n}{n} + \frac{c_2}{n} C_6(\gamma) \log n \Big\} = C_8(\gamma, \eta) \frac{\log n}{n^2} \tag{B.12}$$

which in view of (B.8) and (B.7) completes the proof of part (b).

Appendix C. Relationship between $R_\Phi(\varepsilon)$ and $\underline{T}_\Phi(\varepsilon)$

In this appendix, we prove the claim made in Section 1 about the relation between the upper quantities R_Φ and T_Φ and the lower quantities $\underline{T}_\Phi$ and $\underline{R}_\Phi$. We only carry out the proof for R_Φ; the dual version holds for T_Φ. To simplify the argument, we look at slightly different versions of R_Φ and $\underline{T}_\Phi$, defined as

$$R_\Phi^\circ(\varepsilon) := \sup \big\{ \|f\|_{L^2}^2 : \; f \in B_{\mathcal{H}},\; \|f\|_\Phi^2 < \varepsilon^2 \big\}, \tag{C.1}$$

$$\underline{T}_\Phi^\circ(\delta) := \inf \big\{ \|f\|_\Phi^2 \; : \; f \in B_{\mathcal{H}},\; \|f\|_{L^2}^2 > \delta^2 \big\} \tag{C.2}$$

and prove the following

$$R_\Phi^{\circ\,-1}(\delta) = \underline{T}_\Phi^\circ(\delta) \tag{C.3}$$

where $R_\Phi^{\circ\,-1}(\delta) := \inf\{\varepsilon^2 :\ R_\Phi^\circ(\varepsilon) > \delta^2\}$ is a generalized inverse of R_Φ°. To see (C.3), we note that $R_\Phi(\varepsilon) > \delta^2$ iff there exists $f \in B_{\mathcal{H}}$ such that $\|f\|_\Phi^2 < \varepsilon^2$ and $\|f\|_{L^2}^2 > \delta^2$. But this last statement is equivalent to $\underline{T}_\Phi^\circ(\delta) < \varepsilon^2$. Hence,

$$R_\Phi^{\circ\,-1}(\delta) = \inf\{\varepsilon^2 :\ \underline{T}_\Phi^\circ(\delta) < \varepsilon^2\} \tag{C.4}$$

which proves (C.3).

Using the following lemma, we can use relation (C.3) to convert upper bounds on R_Φ to lower bounds on $\underline{T}_\Phi$.

Lemma 4. *Let $t \mapsto p(t)$ be a nondecreasing function (defined on the real line with values in the extended real line.). Let q be its generalized inverse defined as $q(s) := \inf\{t :\ p(t) > s\}$. Let r be a properly invertible (i.e., one-to-one) function such that $p(t) \le r(t)$, for all t. Then,*

(a) $q(p(t)) \ge t$, for all t,

(b) $q(s) \ge r^{-1}(s)$, for all s.

Proof. Assume (a) does not hold, that is, $\inf\{\alpha :\ p(\alpha) > p(t)\} < t$. Then, there exists α_0 such that $p(\alpha_0) > p(t)$ and $\alpha_0 < t$. But this contradicts $p(t)$ being nondecreasing. For part (b), note that (a) implies $t \le q(p(t)) \le q(r(t))$, since q is nondecreasing by definition. Letting $t := r^{-1}(s)$ and noting that $r(r^{-1}(s)) = s$, by assumption, proves (b). □

Let $p = R_\Phi^\circ$, $q = \underline{T}_\Phi^\circ$ and $r(t) = At + B$ for some constant $A > 0$. Noting that $R_\Phi^\circ \le R_\Phi$ and $\underline{T}_\Phi(\cdot + \gamma) \ge \underline{T}_\Phi^\circ$ for any $\gamma > 0$, we obtain from Lemma 4 and (C.3) that

$$R_\Phi(\varepsilon) \le A\,\varepsilon^2 + B \implies \underline{T}_\Phi(\delta+) \ge \frac{\delta^2}{A} - B, \tag{C.5}$$

where $\underline{T}_\Phi(\delta+)$ denotes the right limit of $\underline{T}_\Phi$ as δ^2. This may be used to translate an upper bound of the form (17) on R_Φ to a corresponding lower bound on $\underline{T}_\Phi$.

Appendix D. The 2×2 subproblem

The following subproblem arises in the proof of Theorem 1.

$$F(\varepsilon^2) := \sup \Big\{ \underbrace{\begin{pmatrix} r & s \end{pmatrix} \begin{pmatrix} u^2 & 0 \\ 0 & v^2 \end{pmatrix} \begin{pmatrix} r \\ s \end{pmatrix}}_{=: x(r,s)} : \ r^2 + s^2 \le 1, \ \underbrace{\begin{pmatrix} r & s \end{pmatrix} \begin{pmatrix} a^2 & 0 \\ 0 & d^2 \end{pmatrix} \begin{pmatrix} r \\ s \end{pmatrix}}_{=: y(r,s)} \le \varepsilon^2 \Big\}, \tag{D.1}$$

where u^2, v^2, a^2 and d^2 are given constants and the optimization is over (r, s). Here, we discuss the solution in some detail; in particular, we provide explicit formulas for $F(\varepsilon^2)$. Without loss of generality assume $u^2 \ge v^2$. Then, it is clear that $F(\varepsilon^2) \le u^2$ and $F(\varepsilon^2) = u^2$ for $\varepsilon^2 \ge u^2$. Thus, we are interested in what happens when $\varepsilon^2 < u^2$.

The problem is easily solved by drawing a picture. Let $x(r, s)$ and $y(r, s)$ be as denoted in the last display. Consider the set

$$\begin{aligned} \mathcal{S} &:= \big\{ \big(x(r,s),\, y(r,s)\big) : \ r^2 + s^2 \le 1 \big\} \\ &= \big\{ r^2(u^2, a^2) + s^2(v^2, d^2) + q^2(0,0) : \ r^2 + s^2 + q^2 = 1 \big\} \\ &= \operatorname{conv} \big\{ (u^2, a^2),\, (v^2, d^2),\, (0,0) \big\}. \end{aligned} \tag{D.2}$$

That is, $\mathcal{S}$ is the convex hull of the three points (u^2, a^2), (v^2, d^2) and the origin $(0, 0)$.

Then, two (or maybe three) different pictures arise depending on whether $a^2 > d^2$ (and whether $d^2 \ge v^2$ or $d^2 < v^2$) or $a^2 \le d^2$; see Fig. D.4. It follows that we have two (or three) different pictures for the function $\varepsilon^2 \mapsto F(\varepsilon^2)$. In particular, for $a^2 > d^2$ and $d^2 < v^2$,

$$F(\varepsilon^2) = v^2 \min \Big\{ \frac{\varepsilon^2}{d^2}, 1 \Big\} + (u^2 - v^2) \max \Big\{ 0, \frac{\varepsilon^2 - d^2}{a^2 - d^2} \Big\}, \tag{D.3}$$

for $a^2 > d^2$ and $d^2 \ge v^2$, $F(\varepsilon^2) = \varepsilon^2$, and for $a^2 \le d^2$,

$$F(\varepsilon^2) = u^2 \min \Big\{ \frac{\varepsilon^2}{a^2}, 1 \Big\}.$$

All the equations above are valid for $\varepsilon^2 \in [0, \sigma_1]$.

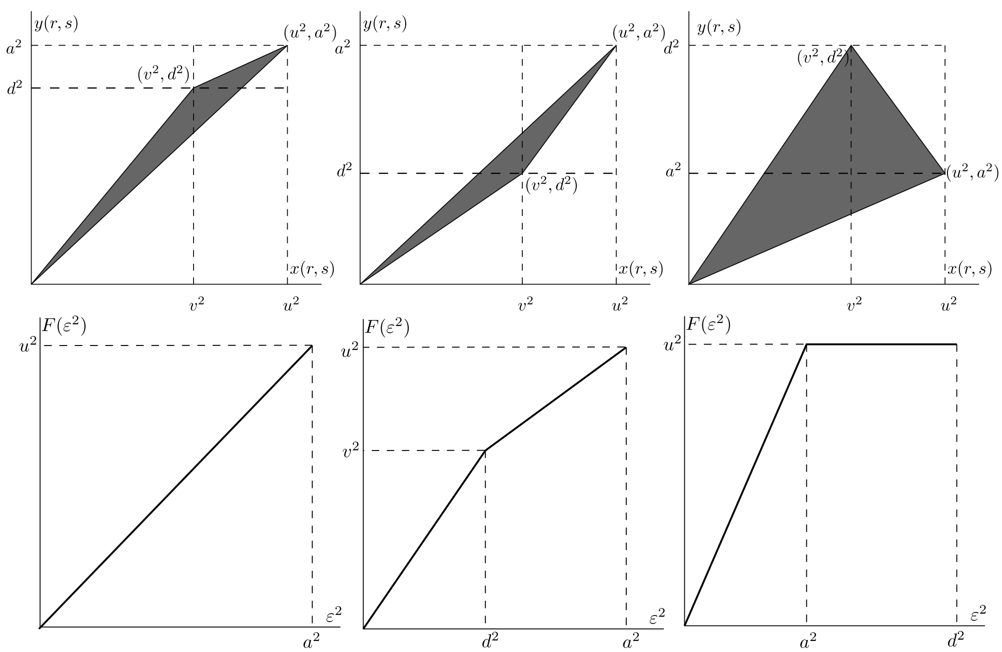


Figure D.4: Top plots illustrate the set $\mathcal{S}$ as defined in (D.2), in various cases. The bottom plots are the corresponding $\varepsilon^2 \mapsto F(\varepsilon^2)$.

Appendix E. Details of the Fourier truncation example

Here we establish the claim that the bound (19) holds with equality. Recall that for the (generalized) Fourier truncation operator $\mathbb{T}_{\psi_1^n}$, we have

$$R_{\mathbb{T}_{\psi_1^n}}(\varepsilon^2) = \sup\Big\{ \sum_{k=1}^{\infty} \sigma_k \alpha_k^2 : \sum_{k=1}^{\infty} \alpha_k^2 \le 1, \sum_{k=1}^{n} \sigma_k \alpha_k^2 \le \varepsilon^2 \Big\}$$

Let $\alpha = (t\xi, s\gamma)$, where $t, s \in \mathbb{R}$, $\xi = (\xi_1, \dots, \xi_n) \in \mathbb{R}^n$, $\gamma = (\gamma_1, \gamma_2 \dots) \in \ell_2$ and $\|\xi\|_2 = 1 = \|\gamma\|_2$. Let $u^2 = u^2(\xi) := \sum_{k=1}^n \sigma_k \xi_k^2$ and $v^2 = v^2(\gamma) := \sum_{k>n} \sigma_k \gamma_k^2$.

Let us fix ξ and γ for now and try to optimize over t and s. That is, we look at

$$G(\varepsilon^2; \xi, \gamma) := \sup\Big\{ t^2 u^2 + s^2 v^2 : t^2 + s^2 \le 1, \ t^2 u^2 \le \varepsilon^2 \Big\}.$$

This is an instance of the 2-by-2 problem (D.1), with $a^2 = u^2$ and $d^2 = 0$. Note that our assumption that $u^2 \ge v^2$ holds in this case, for all ξ and γ, because $\{\sigma_k\}$ is a nonincreasing sequence. Hence, we have, for $\varepsilon^2 \leq \sigma_1$,

$$G(\varepsilon^2; \xi, \gamma) = v^2 + (u^2 - v^2)\frac{\varepsilon^2}{u^2} = v^2(\gamma) + \Big(1 - \frac{v^2(\gamma)}{u^2(\xi)}\Big)\varepsilon^2.$$

Now we can maximize $G(\varepsilon^2; \xi, \gamma)$ over ξ and then γ. Note that G is increasing in u^2. Thus, the maximum is achieved by selecting u^2 to be $\sup_{\|\xi\|_2=1} u^2(\xi) = \sigma_1$. Thus,

$$\sup_{\xi} G(\varepsilon^2; \xi, \gamma) = \Big(1 - \frac{\varepsilon^2}{\sigma_1}\Big)v^2(\gamma) + \varepsilon^2.$$

For $\varepsilon^2 < \sigma_1$, the above is increasing in v^2. Hence the maximum is achieved by setting v^2 to be $\sup_{\|\gamma\|_2=1} v^2(\gamma) = \sigma_{n+1}$. Hence, for $\varepsilon^2 \leq \sigma_1$

$$R_{\mathbb{T}_{\psi_1^n}}(\varepsilon^2) := \sup_{\xi, \gamma} G(\varepsilon^2; \xi, \gamma) = \Big(1 - \frac{\sigma_{n+1}}{\sigma_1}\Big)\varepsilon^2 + \sigma_{n+1}. \tag{E.1}$$

Appendix F. An quadratic inequality

In this appendix, we derive an inequality which will be used in the proof of Theorem 1. Consider a positive semidefinite matrix M (possibly infinite-dimensional) partitioned as

$$M = \begin{pmatrix} A & C \\ C^T & D \end{pmatrix}.$$

Assume that there exists $\rho^2 \in (0, 1)$ and $\kappa^2 > 0$ such that

$$\begin{pmatrix} A & C \\ C^T & (1 - \rho^2)D + \kappa^2 I \end{pmatrix} \succeq 0. \tag{F.1}$$

Let (x, y) be a vector partitioned to match the block structure of M. Then we have the following.

Lemma 5. *Under (F.1), for all x and y,*

$$x^T Ax + 2x^T Cy + y^T Dy \ \geq \ \rho^2 x^T Ax - \frac{\kappa^2}{1 - \rho^2}\|y\|_2^2. \tag{F.2}$$

Proof. By assumption (F.1), we have

$$\begin{pmatrix} \sqrt{1-\rho^2}\, x^T & \frac{1}{\sqrt{1-\rho^2}}\, y^T \end{pmatrix} \begin{pmatrix} A & C \\ C^T & (1-\rho^2)D + \kappa^2 I \end{pmatrix} \begin{pmatrix} \sqrt{1-\rho^2}\, x \\ \frac{1}{\sqrt{1-\rho^2}}\, y \end{pmatrix} \geq 0. \tag{F.3}$$

□

Writing (F.1) as a perturbation of the original matrix,

$$\begin{pmatrix} A & C \\ C^T & D \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & -\rho^2 D + \kappa^2 I \end{pmatrix} \succeq 0, \tag{F.4}$$

we observe that a sufficient condition for (F.1) to hold is $\rho^2 D \preceq \kappa^2 I$. That is, it is sufficient to have

$$\rho^2 \lambda_{\max}(D) \leq \kappa^2. \tag{F.5}$$

Rewriting (F.1) differently, as

$$\begin{pmatrix} (1-\rho^2)A & 0 \\ 0 & (1-\rho^2)D \end{pmatrix} + \begin{pmatrix} \rho^2 A & C \\ C^T & \kappa^2 I \end{pmatrix} \succeq 0, \tag{F.6}$$

we find another sufficient condition for (F.1), namely, $\rho^2 A - \kappa^{-2} C C^T \succeq 0$. In particular, it is also sufficient to have

$$\kappa^{-2} \lambda_{\max}(CC^T) \leq \rho^2 \lambda_{\min}(A). \tag{F.7}$$

References

[1] R. DeVore, Approximation of functions, in: Proc. Symp. Applied Mathematics, Vol. 36, 1986, pp. 1–20.

[2] A. Pinkus, N-Widths in Approximation Theory (Ergebnisse Der Mathematik Und Ihrer Grenzgebiete 3 Folge), Springer, 1985.

[3] A. Pinkus, N-widths and optimal recovery, in: Proc. Symp. Applied Mathematics, Vol. 36, 1986, pp. 51–66.

[4] S. A. van de Geer, Empirical Processes in M-Estimation, Cambridge University Press, 2000.

[5] F. Riesz, B. Sz.-Nagy, Functional Analysis, Dover Publications, 1990.

[6] D. J. H. Garling, Inequalities: a journey into linear analysis, Cambridge Univ Pr, 2007.

[7] R. V. Gamkrelidze, D. Newton, V. M. Tikhomirov, Analysis: Convex analysis and approximation theory, Birkhäuser, 1990.

[8] R. Ahlswede, A. Winter, Strong converse for identification via quantum channels, IEEE Transactions on Information Theory 48 (3) (2002) 569–579. doi:10.1109/18.985947.

[9] M. Rudelson, Random Vectors in the Isotropic Position,, Journal of Functional Analysis 164 (1) (1999) 60–72. doi:10.1006/jfan.1998.3384.

[10] R. Vershynin, Introduction to the non-asymptotic analysis of random matrices, uRL: http://www-personal.umich.edu/ romanv/papers/non-asymptotic-rmt-plain.pdf.

[11] J. A. Tropp, User-friendly tail bounds for sums of random matricesURL: http://arxiv.org/abs/1004.4389.

[12] A. Wigderson, D. Xiao, Derandomizing the Ahlswede-Winter matrix-valued Chernoff bound using pessimistic estimators, and applications, Theory of Computing 4 (1) (2008) 53–76. doi:10.4086/toc.2008.v004a003.