Sampling Theory, Signal Processing, and Data Analysis (2021) 19:13
https://doi.org/10.1007/s43670-021-00013-3

ORIGINAL ARTICLE

L_2-norm sampling discretization and recovery of functions from RKHS with finite trace

Moritz Moeller[1] · Tino Ullrich[1]

Received: 31 October 2020 / Accepted: 31 May 2021 / Published online: 20 July 2021

Abstract
In this paper we study L_2-norm sampling discretization and sampling recovery of complex-valued functions in RKHS on $D \subset \mathbb{R}^d$ based on random function samples. We only assume the finite trace of the kernel (Hilbert–Schmidt embedding into L_2) and provide several concrete estimates with precise constants for the corresponding worst-case errors. In general, our analysis does not need any additional assumptions and also includes the case of non-Mercer kernels and also non-separable RKHS. The fail probability is controlled and decays polynomially in n, the number of samples. Under the mild additional assumption of separability we observe improved rates of convergence related to the decay of the singular values. Our main tool is a spectral norm concentration inequality for infinite complex random matrices with independent rows complementing earlier results by Rudelson, Mendelson, Pajor, Oliveira and Rauhut.

Keywords Spectral norm concentration · Least squares approximation · Random sampling · Discretization · Marcinkiewicz–Zygmund inequalities

Mathematics Subject Classification 41A25 · 41A60 · 41A63 · 68Q25 · 94A20

1 Introduction

This paper can be seen as a continuation of [11,13]. We study the reconstruction of complex-valued multivariate functions on a domain $D \subset \mathbb{R}^d$ from values at the (randomly sampled) nodes $\mathbf{X} := (\mathbf{x}^1, \ldots, \mathbf{x}^n) \in D^n$ via weighted least squares algorithms. In addition, we are interested in the sampling discretization of the squared L_2-norm of such functions using n random nodes. Both problems recently gained substantial interest, see [11,13,14,25–27,29], and are strongly related as we know from Wasilkowski

Communicated by Deguang Han.

✉ Tino Ullrich
tino.ullrich@mathematik.tu-chemnitz.de

[1] Faculty of Mathematics, TU Chemnitz, 09107 Chemnitz, Germany

[30] and the recent systematic studies by Temlyakov [26] and Gröchenig [9] on L_p-norm discretization. Our main interest is on accurate estimates for worst-case errors depending on the number n of nodes. In this paper, the functions are modeled as elements from some reproducing kernel Hilbert space $H(K)$, which is supposed to be compactly embedded into $L_2(D, \varrho_D)$. Its kernel is a positive definite Hermitian function $K\colon D \times D \to \mathbb{C}$. In the papers [11,13,18] authors mainly restrict to the case of separable RKHS [11,18] or Mercer kernels on compact domains [13] with finite trace property to study the quantity

$$\sup_{\|f\|_{H(K)} \le 1} \int_D |f(\mathbf{x}) - S^m_{\mathbf{X}} f(\mathbf{x})|^2 \, d\varrho_D(\mathbf{x}) \tag{1.1}$$

for some recovery operator $S^m_{\mathbf{X}}$. It computes a best least squares fit $S^m_{\mathbf{X}} f$ to the given data

$$\mathbf{f} = (f(\mathbf{x}^1), \ldots, f(\mathbf{x}^n))^\top$$

from the finite-dimensional space spanned by the first $m-1$ singular vectors $\eta_1(\cdot), \ldots, \eta_{m-1}(\cdot)$ of the embedding

$$\mathrm{Id} : H(K) \to L_2(D, \varrho_D). \tag{1.2}$$

We complement existing results by a refined analysis based on spectral norm concentration of infinite matrices to improve on the constants and the bounds for the failure probability on the one hand. On the other hand, the question remained whether the bounds on (1.1) may be extended to the most general situation where only the finite trace condition is assumed. This setting is not covered by the above mentioned references. In this paper we construct a new (weighted) least squares algorithm for this general situation, which has been first addressed by Wasilkowski, Woźniakowski in [31]. Surprisingly, we were able to improve on the bound in [31, Thm. 1] by obtaining the worst-case bound $o(\sqrt{\log n/n})$ in case of square summable singular values $(\sigma_k)_k$ (finite trace) of the embedding. It seems that, in general, their decay influences the bounds rather weakly (in contrast to the results in [11,13,18]).

In addition to the general sampling recovery problem we study the discretization of L_2-integral norms in reproducing kernel Hilbert spaces $H(K)$ where only random information is used. To be more precise, we provide bounds for the following L_2-worst-case discretization errors

$$\sup_{\|f\|_{H(K)} \le 1} \left| \int_D |f(\mathbf{x})|^2 d\varrho_D(\mathbf{x}) - \frac{1}{n} \sum_{i=1}^n |f(\mathbf{x}^i)|^2 \right|. \tag{1.3}$$

This quantity controls the simultaneous discretization of the squared $L_2(D, \varrho_D)$-norms of all functions from $H(K)$. For finite-dimensional spaces we speak of Marcinkiewicz–Zygmund inequalities, a classical topic which also gained a lot of interest in recent years, see Temlyakov [26] and the references therein. Let us emphasize that both problems (sampling recovery and discretization) are strongly related. It has been shown by Wasilkowski [30] that the recovery of the norm $\|\cdot\|$ of a function from a function class is equally difficult as the recovery of the function in that norm using linear information. In other words, if we have a good sampling recovery operator Sf in $L_2(D, \varrho_D)$ we may construct an equally good recovery for the norm of f by simply taking $\|Sf\|_{L_2(D,\varrho_D)}$ as approximant. This, however, is a simple consequence of the triangle inequality. Wasilkowski shows even more, namely that optimal information for the recovery problem is nearly optimal for the "norm-recovery" problem. However, let us emphasize that we recover the square of the norm in (1.3) (rather than the norm itself). It has been observed by Temlyakov in [25] that this indeed makes a difference if we assume a certain algebra property for point-wise multiplication, namely $\|fg\|_{H(K)} \leq c\|f\|_{H(K)} \cdot \|g\|_{H(K)}$ which is for instance present for mixed Sobolev spaces with smoothness $s > 1/2$. Taking into account that in this framework optimal quadrature behaves asymptotically better than sampling recovery (the improvement happens in the log), see [8, Chap. 5, 9] and the references therein, we see that Wasilkowski's result does not hold true for this slightly modified framework. In fact, the worst-case error (1.3) may behave much better than the corresponding optimal sampling recovery error. In contrast to that we use random information here, i.e, nodes which are randomly drawn according to the natural (probability measure) ϱ_D or some related measure and aim for results with high probability. As stated below we obtain a less good asymptotic error behavior for the classical discretization operator in (1.3) compared to the (non-squared) sampling recovery error in (1.1). However, we are able to control the dependence on the parameters and the failure probability rather explicit as (1.4), (1.6), Theorems 1.2 and 1.3 show.

Major parts of the analysis in this paper are based on the following concentration inequality for sums of complex self adjoint (infinite) random matrices.

Theorem 1.1 (Section 3) *Let $\mathbf{y}^i$, $i = 1 \dots n$, be i.i.d random sequences from the complex ℓ_2. Let further $n \geq 3$ and $M > 0$ such that $\|\mathbf{y}^i\|_2 \leq M$ almost surely for all $i = 1 \dots n$. We further put $\Lambda := \mathbb{E}\mathbf{y}^i \otimes \mathbf{y}^i$ and assume that $\|\Lambda\|_{2\to 2} \leq 1$. Then we have for $0 < t \leq 1$*

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}^i \otimes \mathbf{y}^i - \Lambda\right\|_{2\to 2} \geq t\right) \leq 2^{\frac{3}{4}} n \exp\left(-\frac{t^2 n}{21 M^2}\right).$$

Finite-dimensional results of this type are given by Rudelson [23], Tropp [28], Oliveira [20], Rauhut [22] and others. Mendelson and Pajor [16] were the first who addressed the infinite-dimensional case of real matrices as well, see Remark 3.1. The technique used has been introduced by Buchholz [2,3] and further developed by Rauhut for the purpose of analyzing RIP matrices based on complex bounded orthonormal systems (see [22] and the references therein). It is based on an operator version of the non-commutative Khintchine inequality [2,3] together with Talagrand's symmetrization technique.

As a direct consequence of Theorem 1.1 we obtain for separable $H(K)$ and a probability measure ϱ_D always

$$\begin{aligned}&\mathbb{P}\left(\sup_{\|f\|_{H(K)}\le 1}\left|\int_D |f(\mathbf{x})|^2\,d\varrho_D(\mathbf{x})-\frac{1}{n}\sum_{i=1}^n |f(\mathbf{x}^i)|^2\right|>t\|\mathrm{Id}\|_{K,2}^2\right)\\&\le 2^{3/4}n\exp\left(-\frac{t^2n\|\mathrm{Id}\|_{K,2}^2}{21\|K\|_\infty^2}\right)\end{aligned}\qquad(1.4)$$

if the kernel is bounded, i.e., $\|K\|_\infty:=\sup_{\mathbf{x}\in D}\sqrt{K(\mathbf{x},\mathbf{x})}<\infty$ (uniform boundedness). This condition is equivalent to the fact that the embedding of $H(K)$ into ℓ_∞ is continuous and has norm less or equal a finite number M (commonly called M-boundedness). The measure ϱ_D is supposed to be a probability measure and $\mathbf{x}^i$, $i=1,\ldots,n$ are drawn independently at random according to ϱ_D. Note, that this problem is related to classical uniform bounds on the "defect function" in learning theory with respect to M-bounded function classes, see, e.g., [4,6]. There, bounds for (1.3) are usually given in terms of covering (or entropy) numbers of the unit ball of $H(K)$ in ℓ_∞, see [4,12]. Here we consider situations where we neither have such information nor an embedding into ℓ_∞. Choosing t appropriately (see Theorem 6.1), the worst-case discretization error may be bounded as $\mathcal{O}(\sqrt{(\log n)/n})$ with high probability. To get rid of the uniform boundedness condition of the function class we may work with the weaker finite trace condition

$$\operatorname{tr}(K):=\int_D K(\mathbf{x},\mathbf{x})d\varrho_D(\mathbf{x})<\infty\qquad(1.5)$$

and prove a similar error bound for a slightly modified discretization operator when sampling the nodes $\mathbf{x}^i$ independently according to the modified measure $\nu(\mathbf{x})d\varrho_D(\mathbf{x})$ with $\nu(\mathbf{x}):=K(\mathbf{x},\mathbf{x})/\operatorname{tr}(K)$. One only has to replace $\|K\|_\infty^2$ by $\operatorname{tr}(K)$ in the right-hand side of (1.4). In other words, we have

$$\sup_{\|f\|_{H(K)}\le 1}\left|\int_D |f(\mathbf{x})|^2\,d\varrho_D(\mathbf{x})-\frac{1}{n}\sum_{i=1}^n\frac{|f(\mathbf{x}^i)|^2}{\nu(\mathbf{x}^i)}\right|\le\sqrt{21\operatorname{tr}(K)\|\mathrm{Id}\|^2 r\frac{\log n}{n}}\qquad(1.6)$$

with probability exceeding $1-2n^{1-r}$ for large enough n, see Theorem 6.3. This means that the success probability tends to 1 rather quickly as the number of samples increases.

As for the sampling recovery problem we start with a result in the most general situation. A modification of the recovery operator $\widetilde{S}_{\mathbf{X}}^m$ from [11,13], see Algorithm 1 below, has been used to study the situation which is left as an open problem in [11]. The result reads as follows.

Theorem 1.2 (Section 7) *Let $H(K)$ be a reproducing kernel Hilbert space on a subset $D\subset\mathbb{R}^d$ with a positive definite Hermitian kernel $K(\cdot,\cdot)$ such that the finite trace property* (1.5) *holds true. Let $r>1$ and $m,n\in\mathbb{N}$, $n\ge 3$, where m is chosen* *according to* (1.7)*. Drawing* $\mathbf{X} = (\mathbf{x}^1, \dots, \mathbf{x}^n)$ *at random according to the product measure* $(\varrho_m(\mathbf{x})d\varrho_D(\mathbf{x}))^n$ *with the density defined in* (7.1)*, we have*

$$\sup_{\|f\|_{H(K)}\le 1} \|f - \widetilde{S}^m_{\mathbf{X}} f\|^2_{L_2(D,\varrho_D)} \le 441 \max\left\{\sigma_m^2, \frac{r\log n}{n}\sum_{j=m}^{\infty}\sigma_j^2, \frac{\mathrm{tr}_0(K)}{n}\right\}$$

with probability at least $1-\eta n^{1-r}$ *where* $\eta = 2^{\frac{3}{4}}+1$*.* $\widetilde{S}^m_{\mathbf{X}}$ *is the least squares operator from Algorithm* 1 *together with* (7.1) *and* $\mathrm{tr}_0(K)$ *is defined in* (4.6)*.*

In fact, we recover all $f \in H(K)$ from sampled values at $\mathbf{X} = (\mathbf{x}^1, \dots, \mathbf{x}^n)$ simultaneously with probability larger than $1 - 3n^{-r}$ by only assuming that the kernel $K(\cdot, \cdot)$ has finite trace (1.5). Note that this result improves on a result by Wasilkowski and Woźniakowski [31], where only the finite trace is required, see also Novak and Woźniakowski [19, Thm. 26.10]. The authors proved (roughly speaking) a rate of $n^{-1/(2+p)}$ for the worst-case error with respect to standard information if the sequence of singular numbers is p-summable for $p \le 2$. We refer to Sect. 7 for further explanation.

In order to define the recovery operator $\widetilde{S}^m_{\mathbf{X}}$ and the sampling density $\varrho_m(\mathbf{x})$ we need to incorporate spectral properties of the embedding (1.2), namely also the left and right singular functions $(e_k)_k \subset H(K)$ and $(\eta_k)_k \subset L_2(D, \varrho_D)$ ordered according to their importance (size of the corresponding singular number). Both systems are orthonormal in the respective spaces related by $e_k = \sigma_k \eta_k$.

The above result can be improved essentially if we assume that $H(K)$ is separable. This is for instance the case if K is a Mercer kernel, i.e., continuous on a bounded and compact domain D. However, assuming only separability of $H(K)$ also includes the situation of continuous kernels on unbounded domains D, even $D = \mathbb{R}^d$. The following result already improves on the result given in [11,13] in several directions. The theorem works under less restrictive assumptions, the constants are improved and, last but not least, the failure probability decays polynomially in n. We would like to point that, while preparing this manuscript, Ullrich [29] proved a version of the next theorem with stronger requirements and different constants based on Oliveira's concentration result (see Remark 3.9). The following theorem is a reformulation of Theorem 5.2 in Sect. 5.

Theorem 1.3 (Section 5) *Let K be a positive definite Hermitian kernel such that $H(K)$ is separable and the finite trace condition* (1.5) *holds true. With the notation from above we have for $n \in \mathbb{N}$ and*

$$m := \left\lfloor \frac{n}{14r\log n} \right\rfloor \tag{1.7}$$

the bound

$$\mathbb{P}\left(\sup_{\|f\|_{H(K)}\le 1} \|f - \widetilde{S}^m_{\mathbf{X}} f\|^2_{L_2(D,\varrho_D)} \le \frac{15}{m}\sum_{j=\lfloor m/2\rfloor}^{\infty}\sigma_j^2\right) \ge 1 - 3n^{1-r},$$

where $\mathbf{X} = (\mathbf{x}^1, \dots, \mathbf{x}^n)$ *is sampled according to the product measure* $(\varrho_m(\mathbf{x}) d\varrho_D(\mathbf{x}))^n$ (*see* (5.3)) *and the operator* $\widetilde{S}^m_{\mathbf{X}}$ *is defined in Algorithm* 1.

We would like to emphasize that the operator $\widetilde{S}^m_{\mathbf{X}}$ uses $n \asymp m \log m$ samples of its argument. Based on this result it has been recently shown by the second named author (and coauthors, see [18]) that there exists a sampling operator $\widetilde{S}^m_J$ using only $\mathcal{O}(m)$ samples yielding the bound

$$\sup_{\|f\|_{H(K)} \le 1} \|f - \widetilde{S}^m_J f\|^2_{L_2(D,\varrho_D)} \le \frac{C \log(m)}{m} \sum_{j=\lfloor cm \rfloor}^{\infty} \sigma_j^2$$

with universal and specified constants $C, c > 0$. However, for this improvement one has to sacrifice the high success probability.

Notation As usual $\mathbb{N}$ denotes the natural numbers, $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$, $\mathbb{Z}$ denotes the integers, $\mathbb{R}$ the real numbers and $\mathbb{R}_+$ the non-negative real numbers and $\mathbb{C}$ the complex numbers. If not indicated otherwise $\log(\cdot)$ denotes the natural logarithm of its argument. $\mathbb{C}^n$ denotes the complex n-space, whereas $\mathbb{C}^{m \times n}$ denotes the set of all $m \times n$-matrices $\mathbf{L}$ with complex entries. Vectors and matrices are usually typesetted boldface with $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$. The matrix $\mathbf{L}^*$ denotes the adjoint matrix. The spectral norm of matrices $\mathbf{L}$ is denoted by $\|\mathbf{L}\|$ or $\|\mathbf{L}\|_{2 \to 2}$. For a complex (column) vector $\mathbf{y} \in \mathbb{C}^n$ (or ℓ_2) we will often use the tensor notation for the matrix

$$\mathbf{y} \otimes \mathbf{y} := \mathbf{y} \cdot \mathbf{y}^* = \mathbf{y} \cdot \overline{\mathbf{y}}^\top \in \mathbb{C}^{n \times n} \text{ (or } \mathbb{C}^{\mathbb{N} \times \mathbb{N}}).$$

For $0 < p \le \infty$ and $\mathbf{x} \in \mathbb{C}^n$ we denote $\|\mathbf{x}\|_p := (\sum_{i=1}^n |x_i|^p)^{1/p}$ with the usual modification in the case $p = \infty$ or $\mathbf{x}$ being an infinite sequence. Operator norms for $T : H(K) \to L_2$ will be denoted with $\|T\|_{K,2}$. As usual we will denote with $\mathbb{E}X$ the expectation of a random variable X on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Given a measurable subset $D \subset \mathbb{R}^d$ and a measure ϱ we denote with $L_2(D, \varrho)$ the space of all square integrable complex-valued functions (equivalence classes) on D with $\int_D |f(\mathbf{x})|^2 \, d\varrho(\mathbf{x}) < \infty$. We will often use $\Omega = D^n$ as probability space with the product measure $\mathbb{P} = d\varrho^n$ if ϱ is a probability measure itself. We sometimes use the notation $f = \mathcal{O}(g)$ for positive functions f, g, which means that there is a constant $c > 0$ with $f(t) \le cg(t)$. In addition we say $f = o(g)$ if $f(t)/g(t) \to 0$ as $t \to \infty$.

2 Concentration results for sums of random matrices

Let us begin with concentration inequalities for the spectral norm of sums of complex rank-1 matrices. Such matrices appear as $\mathbf{L}^*\mathbf{L}$ when studying least squares solutions of over-determined linear systems

$$\mathbf{L} \cdot \mathbf{c} = \mathbf{f},$$

where $\mathbf{L} \in \mathbb{C}^{n \times m}$ is a matrix with $n > m$, $\mathbf{f} \in \mathbb{C}^n$ and $\mathbf{c} \in \mathbb{C}^{m-1}$. It is well-known that the above system may not have a solution. However, we can ask for the vector $\mathbf{c}$

which minimizes the residual $\|\mathbf{f}-\mathbf{L}\cdot\mathbf{c}\|_2$. Multiplying the system with $\mathbf{L}^*$ gives

$$\mathbf{L}^*\mathbf{L}\cdot\mathbf{c}=\mathbf{L}^*\cdot\mathbf{f}$$

which is called the system of normal equations. If $\mathbf{L}$ has full rank then the unique solution of the least squares problem is given by

$$\mathbf{c}=(\mathbf{L}^*\mathbf{L})^{-1}\mathbf{L}^*\cdot\mathbf{f}.$$

For function recovery and discretization problems we will use the following matrix

$$\mathbf{L}_m:=\begin{pmatrix}\eta_1(\mathbf{x}^1) & \eta_2(\mathbf{x}^1) & \cdots & \eta_{m-1}(\mathbf{x}^1)\\ \vdots & \vdots & & \vdots\\ \eta_1(\mathbf{x}^n) & \eta_2(\mathbf{x}^n) & \cdots & \eta_{m-1}(\mathbf{x}^n)\end{pmatrix}=\begin{pmatrix}\mathbf{y}^{1\top}\\ \vdots\\ \mathbf{y}^{n\top}\end{pmatrix}\quad\text{and}\quad\mathbf{f}=\begin{pmatrix}f(\mathbf{x}^1)\\ \vdots\\ f(\mathbf{x}^n)\end{pmatrix},\tag{2.1}$$

for $\mathbf{X}=(\mathbf{x}^1,\ldots,\mathbf{x}^n)\in D^n$ of distinct sampling nodes and a system of functions $(\eta_k)_{k=1}^{m-1}$. We put $\mathbf{y}^i:=(\eta_1(\mathbf{x}^i),\ldots,\eta_{m-1}(\mathbf{x}^i))^\top$, $i=1,\ldots,n$. The coefficients c_k, $k=1,\ldots,m-1$, of the approximant

$$S^m_{\mathbf{X}}f:=\sum_{k=1}^{m-1}c_k\,\eta_k\tag{2.2}$$

are computed via least squares, see Algorithm 1. Note that the mapping $f\mapsto S^m_{\mathbf{X}}f$ is linear for a fixed set of sampling nodes $\mathbf{X}=(\mathbf{x}^1,\ldots,\mathbf{x}^n)\in D^n$.

We start with a concentration inequality for the spectral norm of a matrix of type (2.1). It turns out that in certain situations the complex matrix $\mathbf{L}_m:=\mathbf{L}_m(\mathbf{X})\in\mathbb{C}^{n\times(m-1)}$ has full rank with high probability, where $\mathbf{X}=(\mathbf{x}^1,\ldots,\mathbf{x}^n)$ is drawn at random from D^n according to a product measure $\mathbb{P}=d\varrho^n$. We will find that the eigenvalues of

$$\mathbf{H}_m:=\mathbf{H}_m(\mathbf{X})=\frac{1}{n}\mathbf{L}_m^*\mathbf{L}_m=\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}^i\otimes\mathbf{y}^i\in\mathbb{C}^{(m-1)\times(m-1)},\tag{2.3}$$

are bounded away from zero with high probability if m is small enough compared to n and the functions $\eta_k(\cdot)$ denote an orthonormal system with respect to the measure ϱ from which the nodes in $\mathbf{X}$ are sampled. Let us define the corresponding spectral function

$$N(m):=\sup_{\mathbf{x}\in D}\sum_{k=1}^{m-1}|\eta_k(\mathbf{x})|^2.\tag{2.4}$$

From [28, Theorem 1.1] we get the following result.

Theorem 2.1 (Matrix Chernoff) *For a finite sequence* $(\mathbf{A}_k)$ *of independent, Hermitian, positive semi-definite random matrices with dimension* m *satisfying* $\lambda_{\max}(\mathbf{A}_k) \leq R$ *almost surely it holds*

$$\mathbb{P}\left(\lambda_{\min}\left(\sum_{k=1}^{n} \mathbf{A}_k\right) \leq (1-t)\mu_{\min}\right) \leq m\left(\frac{e^{-t}}{(1-t)^{1-t}}\right)^{\mu_{\min}/R}$$
$$\mathbb{P}\left(\lambda_{\max}\left(\sum_{k=1}^{n} \mathbf{A}_k\right) \geq (1+t)\mu_{\max}\right) \leq m\left(\frac{e^{t}}{(1+t)^{1+t}}\right)^{\mu_{\max}/R}$$

for $t \in [0,1]$ *where* $\mu_{\min} := \lambda_{\min}\left(\sum_{k=1}^{n} \mathbb{E}\mathbf{A}_k\right)$ *and* $\mu_{\max} := \lambda_{\max}\left(\sum_{k=1}^{n} \mathbb{E}\mathbf{A}_k\right)$.

Theorem 2.2 *Let* $n, m \in \mathbb{N}$, $m \geq 2$ *and* $\{\eta_1(\cdot), \eta_2(\cdot), \eta_3(\cdot), \ldots, \eta_{m-1}(\cdot)\}$ *be an orthonormal system in* $L_2(D, \varrho)$. *Let* $\mathbf{H}_m$ *be given as above and* $\mathbf{x}^1, \ldots, \mathbf{x}^n \in D$ *drawn i.i.d. at random according to* $\mathbb{P} = d\varrho$ *we have for* $0 < t < 1$ *that*

$$\mathbb{P}(\lambda_{\min}(\mathbf{H}_m) < 1-t) \leq m \exp\left(-\frac{n \log c_t}{N(m)}\right),$$

as well as

$$\mathbb{P}(\lambda_{\max}(\mathbf{H}_m) > 1+t) \leq m \exp\left(-\frac{n \log d_t}{N(m)}\right),$$

where $c_t := (1-t)^{1-t}e^t$ *and* $d_t := (1+t)^{1+t}e^{-t}$.

Proof We apply Theorem 2.1. To do this we define $\mathbf{A}_i = \frac{1}{n}\mathbf{y}^i \otimes \mathbf{y}^i$. One easily sees that all the matrices $\mathbf{A}_i$ are always positive semi-definite and $\lambda_{\min}\left(\sum_{i=1}^{n} \mathbb{E}\mathbf{A}_i\right) = 1$. We have that

$$\lambda_{\max}(\mathbf{A}_i) = \|\mathbf{y}^i\|^2/n \leq N(m)/n.$$

Plugging this into Theorem 2.1 yields

$$\mathbb{P}(\lambda_{\min}(\mathbf{H}_m) \leq 1-t) \leq m\left[\frac{e^{-t}}{(1-t)^{1-t}}\right]^{n/N(m)} \leq m \exp\left(-\frac{n \log c_t}{N(m)}\right).$$

Theorem 2.3 *For* $n \geq m$ *and* $r > 1$ *the matrix* $\mathbf{H}_m$ *has only eigenvalues greater than* $1/2$ *with probability at least* $1 - n^{1-r}$ *if*

$$N(m) \leq \frac{n}{7r\log n}. \tag{2.5}$$

In particular, we have

$$\|(\mathbf{L}_m^*\mathbf{L}_m)^{-1}\mathbf{L}_m^*\|_{2\to 2} \leq \sqrt{\frac{2}{n}}. \tag{2.6}$$

Proof Choosing $t = 1/2$ and solving for $N(m)$ in the above probability bound (using n^{1-r} on the right-hand side) gives the desired result. Indeed

$$\mathbb{P}(\lambda_{\min}(\mathbf{H}_m) < 1 - t) \leq m \exp\left(-\frac{n \log c_t}{N(m)}\right) \leq n^{1-r}.$$

This gives the following implications (read from bottom to top)

$$\begin{aligned} \log(m) - \log(c_t)\frac{n}{N(m)} &\leq \log n^{1-r} \\ \frac{\log m - \log n^{1-r}}{\log c_t} &\leq \frac{n}{N(m)} \\ N(m) &\leq \frac{n \log c_t}{\log m - \log n^{1-r}} \\ N(m) &\leq \frac{n}{7(\log n - (1-r)\log n)} \\ N(m) &\leq \frac{n}{7\, r \log n}. \end{aligned} \tag{2.7}$$

The bound in (2.6) is a consequence of [11, Proposition 3.1]. □

From [11, Proposition 3.1] we also get a lower bound of $\|(\mathbf{L}_m^*\mathbf{L}_m)^{-1}\mathbf{L}_m^*\|_{2\to 2}$ with high probability.

Corollary 2.4 *Let $\{\eta_1(\cdot), \eta_2(\cdot), \eta_3(\cdot), \ldots\}$ be an orthonormal system in $L_2(D, \varrho)$. Let further $r > 1$ and $m, n \in \mathbb{N}$, $m \geq 2$ such that*

$$N(m) \leq \frac{n}{10\, r \log n}$$

holds. Then the random matrix $\mathbf{L}_m$ from (2.1) satisfies

$$\|(\mathbf{L}_m^*\mathbf{L}_m)^{-1}\mathbf{L}_m^*\|_{2\to 2} \geq \sqrt{\frac{2}{3n}}$$

with probability at least $1 - n^{1-r}$, where the nodes are sampled i.i.d according to ϱ.

3 Norm concentration for infinite matrices

In this section we want to extend some of the results from Sect. 2 to the infinite dimensional framework. We provide a new concentration inequality derived from the non-commutative Khintchine inequality via a bootstrapping argument using a symmetrization result by Ledoux and Talagrand [15] for Rademacher sums of random operators $\mathbf{B}_i = \mathbf{y}^i \otimes \mathbf{y}^i$, where $\mathbf{y}^i$ will denote a complex random infinite ℓ_2-sequence.

Definition 3.1 (*Schatten-p-Norm*) For a compact operator $\mathbf{A} : H \to K$ between complex Hilbert spaces H and K and $1 \leq p \leq \infty$ we define the Schatten-p-norm:

$$\|\mathbf{A}\|_{S_p} = \|\sigma(\mathbf{A})\|_p$$

where $\sigma(\mathbf{A})$ is the vector of singular values of $\mathbf{A}$.

The quantity $\|\cdot\|_{S_p}$ is a norm, see, e.g. [7].

Corollary 3.2 *One easily sees that*

$$\|\mathbf{A}\|_{2\to 2} = \|\mathbf{A}\|_{S_\infty} \leq \|\mathbf{A}\|_{S_p}$$

and that for $\mathbf{A}$ *with rank at most* r *it holds that*

$$\|\mathbf{A}\|_{S_p} \leq r^{1/p} \|\mathbf{A}\|_{2\to 2}$$

for $1 \leq p \leq \infty$.

Definition 3.3 (*Schatten-class*) Let H, K be complex Hilbert spaces and $1 \leq p < \infty$. The p-th Schatten-class is defined as

$$S_p(H, K) := \left\{\mathbf{A} : H \to K, \mathbf{A} \textit{ compact}, \|\mathbf{A}\|_{S_p} < \infty\right\}.$$

Theorem 3.4 (Non-commutative Khintchine inequality, [2,3]) *Let* $n \in \mathbb{N}$ *and* $\mathbf{B}_i$, $i = 1, \dots, m$, *denote operators from* S_{2n}. *Let further* ε_i *denote independent Rademacher variables for* $i = 1 \dots m$. *Then it holds*

$$\mathbb{E}_{\boldsymbol{\varepsilon}} \left\| \sum_{i=1}^m \varepsilon_i \mathbf{B}_i \right\|_{S_{2n}}^{2n} \leq \frac{(2n)!}{2^n n!} \max \left\{ \left\| \left(\sum_{i=1}^m \mathbf{B}_i \mathbf{B}_i^* \right)^{1/2} \right\|_{S_{2n}}^{2n}, \left\| \left(\sum_{i=1}^m \mathbf{B}_i^* \mathbf{B}_i \right)^{1/2} \right\|_{S_{2n}}^{2n} \right\}.$$

Corollary 3.5 (Rudelson's lemma) *Let* $\mathbf{y}^i$ *be a sequence from the complex* ℓ_2 *and* ε_i *independent Rademacher variables for* $i = 1 \dots m$. *Then it holds for* $2 \leq p < \infty$ *that*

$$\left(\mathbb{E}_{\boldsymbol{\varepsilon}} \left\| \sum_{i=1}^m \varepsilon_i \mathbf{y}^i \otimes \mathbf{y}^i \right\|_{2\to 2}^p \right)^{1/p} \leq 2^{3/4p} m^{1/p} \sqrt{p} e^{-\frac{1}{2}} \sqrt{\left\| \sum_{i=1}^m \mathbf{y}^i \otimes \mathbf{y}^i \right\|_{2\to 2}} \max_{i=1\dots m} \|\mathbf{y}^i\|_2.$$

Proof We utilize the non-commutative Khintchine inequality with $\mathbf{B}_i := \mathbf{y}^i \otimes \mathbf{y}^i$ which belong to every S_{2n} since they have (at most) rank 1. Applying literally the same arguments as in [22, Lemma 6.18] we obtain the result (see [17] for details).

We estimate tails of random variables by means of their moments. We will use a well-known relation, see e.g. [22, Proposition 6.5].

Proposition 3.6 (Moments and tails) *Let X be a random variable that for all $p \geq p_0$ satisfies*

$$\left(\mathbb{E}|X|^p\right)^{\frac{1}{p}} \leq \alpha\beta^{\frac{1}{p}} p^{\frac{1}{\gamma}}$$

for some constants $\alpha, \beta, \gamma, p_0 > 0$. Then

$$\mathbb{P}\left(|X| \geq e^{\frac{1}{\gamma}} \alpha u\right) \leq \beta e^{-\frac{u^\gamma}{\gamma}}$$

for all $u \geq p_0^{1/\gamma}$.

From [15, Lemma 6.3] we get the following result.

Proposition 3.7 (Symmetrization, [15]) *Let $F : \mathbb{R}_+ \to \mathbb{R}_+$ be convex. Let $\mathbf{X}_i$, $i = 1, \ldots n$, be independent random variables in a separable Banach space $(B, \|\cdot\|)$ such that $\mathbb{E}F(\|\mathbf{X}_i\|) < \infty$. Let further $\boldsymbol{\varepsilon} = (\varepsilon_i)_{i=1}^n$ be independent Rademacher variables which are also independent of $\mathbf{X}_i$. Then it holds that*

$$\mathbb{E}F\left(\sup_{f\in D}\left|f\left(\sum_{i=1}^n \mathbf{X}_i\right) - \mathbb{E}f\left(\sum_{i=1}^n \mathbf{X}_i\right)\right|\right) \leq \mathbb{E}F\left(2\left\|\sum_{i=1}^n \varepsilon_i \mathbf{X}_i\right\|\right),$$

where D is a countable set of linear functionals with $\|\mathbf{x}\| = \sup_{f\in D}\left|f(\mathbf{x})\right|$ for all $\mathbf{x} \in B$.

Proposition 3.8 *Let $\mathbf{y}^i$, $i = 1 \ldots n$, be i.i.d random sequences from the complex ℓ_2. Let further $n \geq 3$, $r > 1$, $M > 0$ such that $\|\mathbf{y}^i\|_2 \leq M$ for all $i = 1 \ldots n$ almost surely and $\mathbb{E}\mathbf{y}^i \otimes \mathbf{y}^i = \Lambda$ for all $i = 1 \ldots n$. Then*

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{y}^i \otimes \mathbf{y}^i - \Lambda\right\|_{2\to 2} \geq F\right) \leq 2^{\frac{3}{4}} n^{1-r},$$

where $F := \max\left\{\frac{8r\log n}{n} M^2\kappa^2, \|\Lambda\|_{2\to 2}\right\}$ and $\kappa = \frac{1+\sqrt{5}}{2}$.

Proof We use a method as in [22, Theorem 7.3]. For $2 \leq p < \infty$ we put

$$E_p := \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{y}^i \otimes \mathbf{y}^i - \Lambda\right\|_{2\to 2}^p.$$

Since $\sum_{i=1}^n \mathbf{y}^i \otimes \mathbf{y}^i$ has rank (at most) n it is compact. The expectation matrix Λ is a positive semidefinite operator with finite trace since $\|\mathbf{y}^i\|_2 \leq M$ for all $i = 1 \ldots n$ almost surely. This means Λ is a trace class operator and therefore compact. Since $\frac{1}{n}\sum_{i=1}^n \mathbf{y}^i \otimes \mathbf{y}^i - \Lambda$ is compact and the subspace of all compact operators $K(\ell_2, \ell_2)$

from ℓ_2 to ℓ_2 is separable we can choose a countable set D from the dual space of $K(\ell_2, \ell_2)$ as in Proposition 3.7 such that

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{y}^i \otimes \mathbf{y}^i - \Lambda \right\|_{2\to 2} = \sup_{f \in D} \left| f\left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{y}^i \otimes \mathbf{y}^i \right) - f\left(\mathbb{E} \frac{1}{n} \sum_{i=1}^{n} \mathbf{y}^i \otimes \mathbf{y}^i \right) \right|.$$

Since f is a continuous linear functional we get

$$E_p = \mathbb{E}\left(\sup_{f \in D} \left| f\left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{y}^i \otimes \mathbf{y}^i \right) - \mathbb{E} f\left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{y}^i \otimes \mathbf{y}^i \right) \right|^p \right). \tag{3.1}$$

Proposition 3.7 applied to (3.1) with $F(t) = t^p$ together with Rudelson's lemma (Corollary 3.5) for $2 \le p < \infty$ in the form

$$\left(\mathbb{E}_{\boldsymbol{\varepsilon}} \left\| \sum_{i=1}^{n} \varepsilon_i \, \mathbf{y}^i \otimes \mathbf{y}^i \right\|_{2\to 2}^{p} \right)^{1/p} \le 2^{3/4p} n^{1/p} \sqrt{p} e^{-\frac{1}{2}} \sqrt{\left\| \sum_{i=1}^{n} \mathbf{y}^i \otimes \mathbf{y}^i \right\|_{2\to 2}} \max_{i=1\ldots n} \|\mathbf{y}^i\|_2$$

yields

$$\begin{aligned} E_p &\le 2^p \, \mathbb{E}_{\mathbf{y}} \mathbb{E}_{\boldsymbol{\varepsilon}} \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \, \mathbf{y}^i \otimes \mathbf{y}^i \right\|_{2\to 2}^{p} \\ &\le \left(\frac{2}{\sqrt{n}} \right)^p 2^{\frac{3}{4}} n \, p^{\frac{p}{2}} e^{-\frac{p}{2}} \, \mathbb{E}_{\mathbf{y}} \left(\sqrt{\left\| \sum_{i=1}^{n} \mathbf{y}^i \otimes \mathbf{y}^i \right\|_{2\to 2}} \max_{i=1\ldots n} \|\mathbf{y}^i\|_2 \right)^p \\ &\le \left(\frac{2}{\sqrt{n}} \right)^p 2^{\frac{3}{4}} n \, p^{\frac{p}{2}} e^{-\frac{p}{2}} M^p \, \mathbb{E}_{\mathbf{y}} \left(\sqrt{\left\| \sum_{i=1}^{n} \mathbf{y}^i \otimes \mathbf{y}^i \right\|_{2\to 2}} \right)^p \\ &\le \left(\frac{2}{\sqrt{n}} \right)^p 2^{\frac{3}{4}} n \, p^{\frac{p}{2}} e^{-\frac{p}{2}} M^p \left(\sqrt{\mathbb{E}_{\mathbf{y}} \left(\left\| \sum_{i=1}^{n} \mathbf{y}^i \otimes \mathbf{y}^i - \Lambda \right\|_{2\to 2} \right) + \|\Lambda\|_{2\to 2}} \right)^p \end{aligned}$$

because of $\|\mathbf{y}^i\|_2 \le M$ and Hölder's inequality. Using triangle inequality and the fact that $\mathbb{E}\big(\|\mathbf{X}\| + \|\mathbf{Y}\|\big)^p \le \Big(\big(\mathbb{E}\|\mathbf{X}\|^p\big)^{1/p} + \big(\mathbb{E}\|\mathbf{Y}\|^p\big)^{1/p} \Big)^p$ we have

$$E_p \le \left(\frac{2}{\sqrt{n}} \right)^p 2^{\frac{3}{4}} n \, p^{\frac{p}{2}} e^{-\frac{p}{2}} M^p \left(\sqrt{\mathbb{E}_{\mathbf{y}} \left(\left\| \sum_{i=1}^{n} \mathbf{y}^i \otimes \mathbf{y}^i - \Lambda \right\|_{2\to 2}^{p} \right)^{1/p} + \|\Lambda\|_{2\to 2}} \right).$$

Setting $D_{p,n,M} := \frac{2}{\sqrt{n}} 2^{\frac{3}{4p}} M p^{\frac{1}{2}} n^{\frac{1}{p}} e^{-\frac{1}{2}}$ yields

$$E_p^{1/p} \le D_{p,n,M} \sqrt{E_p^{1/p} + F}, \tag{3.2}$$

where $F \ge \|\Lambda\|_{2\to 2}$ will be chosen later. Solving this regarding $E_p^{1/p}$ gives

$$E_p^{1/p} \le \frac{D_{p,n,M}^2}{2} + \sqrt{D_{p,n,M}^2 \cdot F + \frac{D_{p,n,M}^4}{4}}.$$

We now consider the random variable $\min\left\{F, \frac{1}{n}\left\| \sum_{i=1}^n \mathbf{y}^i \otimes \mathbf{y}^i - \Lambda \right\|_{2\to 2}\right\}$. Obviously

$$\left(\mathbb{E} \min\left\{F, \left\| \frac{1}{n}\sum_{i=1}^n \mathbf{y}^i \otimes \mathbf{y}^i - \Lambda \right\|_{2\to 2}\right\}^p\right)^{1/p} \le \min\{F, E_p^{1/p}\}.$$

In the case of $D_{p,n,M}^2 \le F$ we get

$$\min\{F, E_p^{1/p}\} \le E_p^{1/p} \le D_{p,n,M}\sqrt{F}\left(\frac{1+\sqrt{5}}{2}\right) =: D_{p,n,M}\sqrt{F}\,\kappa$$

and otherwise

$$\min\{F, E_p^{1/p}\} \le F \le \ D_{p,n,M}\sqrt{F} \le D_{p,n,M}\sqrt{F}\,\kappa.$$

This yields

$$\left(\mathbb{E} \min\left\{F, \left\| \frac{1}{n}\sum_{i=1}^n \mathbf{y}^i \otimes \mathbf{y}^i - \Lambda \right\|_{2\to 2}\right\}^p\right)^{1/p} \le \kappa D_{p,n,M}\sqrt{F}.$$

Using Proposition 3.6 we get that

$$\mathbb{P}\left(\min\left\{F, \left\| \frac{1}{n}\sum_{i=1}^n \mathbf{y}^i \otimes \mathbf{y}^i - \Lambda \right\|_{2\to 2}\right\} \ge \frac{2}{\sqrt{n}} M\kappa\sqrt{F}u\right) \le 2^{\frac{3}{4}}\, n \exp\left(\frac{-u^2}{2}\right) \tag{3.3}$$

for all $u \ge \sqrt{2}$. Now we choose $u = \sqrt{2r\log n}$ with $r > 1$ and $n \ge 3$.This gives

$$\mathbb{P}\left(\min\left\{F, \left\| \frac{1}{n}\sum_{i=1}^n \mathbf{y}^i \otimes \mathbf{y}^i - \Lambda \right\|_{2\to 2}\right\} \ge \frac{2}{\sqrt{n}} M\kappa\sqrt{F}\sqrt{2r\log n}\right) \le 2^{\frac{3}{4}}\, n^{1-r}.$$

In case $F \geq \frac{2}{\sqrt{n}} M\kappa\sqrt{F}u$ we can avoid the minimum on the left-hand side. It clearly holds

$$F \geq \frac{2}{\sqrt{n}} M\kappa\sqrt{F}u = \frac{2}{\sqrt{n}} M\kappa\sqrt{F}\sqrt{2r\log n},$$

and hence

$$\sqrt{F} \geq \frac{2}{\sqrt{n}} M\kappa\sqrt{2r\log n}.$$

The latter is satisfied if

$$F := \max\left\{\frac{8r\log n}{n} M^2\kappa^2, \|\Lambda\|_{2\to 2}\right\}. \tag{3.4}$$

This yields

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n} \mathbf{y}^i \otimes \mathbf{y}^i - \Lambda\right\|_{2\to 2} \geq F\right) \leq 2^{\frac{3}{4}}\, n^{1-r}.$$

□

Using this result we are now able to proof our main concentration inequality.

Proof of Theorem 1.1

Proof Let us return to (3.2) in the above proof. Since $\|\mathbf{\Lambda}\|_{2\to 2} \leq 1$ we get as a consequence for $0 < t \leq 1$

$$E_p^{1/p} \leq \tilde{D}_{p,n,M}\sqrt{E_p^{1/p} + t} \tag{3.5}$$

with

$$\tilde{D}_{p,n,M} := \frac{1}{\sqrt{t}}\frac{2}{\sqrt{n}} 2^{\frac{3}{4p}} M p^{\frac{1}{2}} n^{\frac{1}{p}} e^{-\frac{1}{2}}.$$

We continue in the proof as above using $\tilde{D}_{p,n,M}$ instead of $D_{p,n,M}$ and without replacing u by $\sqrt{2r\log n}$ in (3.3). With the same argumentation as above we get rid of the minimum and obtain this time

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n} \mathbf{y}^i \otimes \mathbf{y}^i - \Lambda\right\|_{2\to 2} \geq F\right) \leq 2^{\frac{3}{4}} n \exp(-u^2/2) \tag{3.6}$$

for $F \geq \max\{4M^2u^2\kappa^2/(nt), t\}$. The maximum is no longer necessary if we choose $u^2 := t^2 n/(4M^2\kappa^2)$. Plugging this choice into (3.6) and noting that $8\kappa^2 \leq 21$ gives the desired bound. □

Remark 3.9 The first result of this type is due to Rudelson [23] for vectors from $\mathbb{R}^n$. Complex versions were proved by Rauhut [22] and Oliveira [20, Lem. 1]. Note that the result stated by Oliveira (Lemma 1) contains a small incorrectness in the probability bound. A corrected version has been stated in [11, Prop. 4.1]. In his paper Oliveira also comments on the infinite dimensional complex situation where $m = \infty$ but does not give a full proof. The proof method is different from ours. Note also that in [16, Cor. 2.6] Mendelson and Pajor give a concentration result for the infinite dimensional case of real vectors. Let us finally mention that a version of our Theorem 1.3 (in the next section) under more restrictive assumptions has been recently proved by Ullrich [29] based on Oliveira's concentration result.

4 Reproducing kernel Hilbert spaces

We will work in the framework of reproducing kernel Hilbert spaces. The relevant theoretical background can be found in [1, Chapt. 1] and [4, Chapt. 4]. The papers [10,24] are also of particular relevance for the subject of this paper.

Let $L_2(D, \varrho_D)$ be the space of complex-valued square-integrable functions with respect to ϱ_D. Here $D \subset \mathbb{R}^d$ is an arbitrary subset and ϱ_D a measure on D. We further consider a reproducing kernel Hilbert space $H(K)$ with a Hermitian positive definite kernel $K(\cdot, \cdot)$ on $D \times D$. The crucial property of reproducing kernel Hilbert spaces is the fact that Dirac functionals are continuous, or equivalently, the reproducing property holds

$$f(\mathbf{x}) = \langle f, K(\cdot, \mathbf{x})\rangle_{H(K)}$$

for all $\mathbf{x} \in D$. It ensures that point evaluations are continuous functionals on $H(K)$. We will use the notation from [4, Chapt. 4]. In the framework of this paper, the finite trace of the kernel

$$\mathrm{tr}(K) := \|K\|_2^2 = \int_D K(\mathbf{x}, \mathbf{x}) d\varrho_D(\mathbf{x}) < \infty \tag{4.1}$$

or its boundedness

$$\|K\|_\infty := \sup_{\mathbf{x}\in D} \sqrt{K(\mathbf{x}, \mathbf{x})} < \infty \tag{4.2}$$

is assumed. The boundedness of K implies that $H(K)$ is continuously embedded into $\ell_\infty(D)$, i.e.,

$$\|f\|_{\ell_\infty(D)} \leq \|K\|_\infty \cdot \|f\|_{H(K)}. \tag{4.3}$$

With $\ell_\infty(D)$ we denote the set of bounded functions on D and with $\|\cdot\|_{\ell_\infty(D)}$ the supremum norm. Note, that we do not need the measure ϱ_D for this embedding. In fact, here we mean "boundedness" in the strong sense (in contrast to essential boundedness w.r.t. the measure ϱ_D). The embedding operator

$$\mathrm{Id} : H(K) \to L_2(D, \varrho_D) \tag{4.4}$$

is Hilbert–Schmidt under the finite trace condition (4.1), see [10], [24, Lemma 2.3], which we always assume from now on. We additionally assume that $H(K)$ is at least infinite dimensional. Let us denote the (at most) countable system of strictly positive eigenvalues $(\lambda_j)_{j\in\mathbb{N}}$ of $W_{K,\varrho_D} = \mathrm{Id}^*_{K,\varrho_D} \circ \mathrm{Id}_{K,\varrho_D}$ arranged in non-increasing order, i.e.,

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots > 0.$$

We will also need the left and right singular vectors $(e_k)_k \subset H(K)$ and $(\eta_k)_k \subset L_2(D, \varrho_D)$ which both represent orthonormal systems in the respective spaces related by $e_k = \sigma_k \eta_k$ with $\lambda_k = \sigma_k^2$ for $k \in \mathbb{N}$. We would like to emphasize that the embedding (4.3) is not necessarily injective. In other words, for certain kernels there might be a also a nontrivial nullspace of the embedding Id in (4.4). Therefore, the system $(e_k)_k$ from above is not necessarily a basis in $H(K)$. It would be a basis under additional restrictions, e.g., the kernel $K(\cdot,\cdot)$ is continuous and bounded (Mercer kernel). Based on this observation we will decompose the kernel $K(\cdot,\cdot)$ as follows

$$\begin{aligned} K(\mathbf{x},\mathbf{y}) &= K^0(\mathbf{x},\mathbf{y}) + K^1(\mathbf{x},\mathbf{y}) \\ &:= \left(K(\mathbf{x},\mathbf{y}) - \sum_{k=1}^{\infty} e_k(\mathbf{x})\overline{e_k(\mathbf{y})}\right) + \sum_{k=1}^{\infty} e_k(\mathbf{x})\overline{e_k(\mathbf{y})}. \end{aligned} \tag{4.5}$$

By Bessel's inequality we get that

$$\mathrm{tr}_0(K) := \mathrm{tr}(K^0) = \int_D K(\mathbf{x},\mathbf{x})d\,\varrho_D(\mathbf{x}) - \sum_{k=1}^{\infty} \lambda_k \geq 0. \tag{4.6}$$

It is shown in [10], [24, Lemma 2.3] that if $\mathrm{tr}(K) < \infty$ and $H(K)$ is separable then $\mathrm{tr}_0(K) = 0$. As we will see below, it will make a big difference if $\mathrm{tr}_0(K)$ vanishes or not. The second case is only apparent if $H(K)$ is non-separable. In other words, if $H(K)$ is separable the function $K^0(\mathbf{x},\mathbf{x}) := K(\mathbf{x},\mathbf{x}) - \sum_{k=1}^{\infty} |e_k(\mathbf{x})|^2$ is zero almost everywhere with respect to the measure ϱ_D. Let us finally define the two crucial quantities

$$N(m) := \sup_{\mathbf{x}\in D} \sum_{k=1}^{m-1} |\eta_k(\mathbf{x})|^2 \tag{4.7}$$

and

$$T(m) := \sup_{\mathbf{x}\in D} \sum_{k=m}^{\infty} |e_k(\mathbf{x})|^2. \tag{4.8}$$

The first one is often called "spectral function", see [9] and the references therein.

5 Sampling recovery guarantees for separable RKHS

In this section we deal with the case that $H(K)$ is a separable Hilbert space on a subset $D \subset \mathbb{R}^d$ which is compactly embedded in $L_2(D, \varrho_D)$ for a given measure ϱ_D. The first Theorem gives a result in a more restrictive situation, namely that ϱ_D is a probability measure and the kernel is bounded. In the second theorem we sample with respect to the probability density function ϱ_m defined below in (5.3) and invented by Krieg and Ullrich [13,14]. We use the same proof strategy as in [11]. Here we do not apply Rudelson's bound [23] on the expectation. We rather use the concentration inequality proved in Proposition 3.8. This leads to a polynomial decaying failure probability, see also [29].

Theorem 5.1 *Let $H(K)$ be a separable reproducing kernel Hilbert space on a set $D \subset \mathbb{R}^d$ with a positive definite kernel $K(\cdot, \cdot)$ such that $\sup_{\mathbf{x} \in D} K(\mathbf{x}, \mathbf{x}) < \infty$. We denote with $(\sigma_j)_{j \in \mathbb{N}}$ the non-increasing sequence of singular numbers of the embedding $\mathrm{Id} : H(K) \to L_2(D, \varrho_D)$ for a probability measure ϱ_D. Let further $r > 1$ and $m, n \in \mathbb{N}$, $n \geq 3$ where $m \geq 2$ is chosen such that*

$$N(m) \leq \frac{n}{7r \log n} \tag{5.1}$$

holds. Drawing $\mathbf{X} = (\mathbf{x}^1, \ldots, \mathbf{x}^n)$ according to the product measure $d\varrho_D^n$, we have

$$\mathbb{P}\left(\sup_{\|f\|_{H(K)} \leq 1} \|f - S_{\mathbf{X}}^m f\|_{L_2(D, \varrho_D)}^2 \leq 5 \max\left\{ \sigma_m^2, \frac{8r \log n}{n} T(m) \kappa^2 \right\} \right) \geq 1 - \eta n^{1-r},$$

where $\eta = 2^{\frac{3}{4}} + 1$, $\kappa = \frac{1+\sqrt{5}}{2}$ and $N(m)$, $T(m)$ are defined in (4.7), (4.8).

Proof We define the events

$$A := \left\{ \mathbf{X} \in D^n \ : \ \frac{1}{n} \left\| \mathbf{\Phi}_m^* \mathbf{\Phi}_m \right\|_{2 \to 2} \leq F + \sigma_m^2 \right\},$$
$$B := \left\{ \mathbf{X} \in D^n \ : \ \frac{1}{2} \leq \lambda_i(\mathbf{H}_m), \ i = 1 \ldots m \right\},$$

where F appears in (3.4) and $\mathbf{H}_m$ in (2.3). The operator $\mathbf{\Phi}_m$ is given by

$$\mathbf{\Phi}_m : \ell_2 \to \mathbb{R}^n, \quad \mathbf{z} \mapsto \begin{pmatrix} \langle \mathbf{z}, \mathbf{y}^1 \rangle \\ \vdots \\ \langle \mathbf{z}, \mathbf{y}^n \rangle \end{pmatrix}$$

with $\mathbf{y}^i = (e_m(\mathbf{x}^i), e_{m+1}(\mathbf{x}^i) \ldots)^\top$ for all $i = 1 \ldots n$. Hence we may choose F as

$$F := \max\left\{ \frac{8r \log n}{n} T(m) \kappa^2, \sigma_m^2 \right\}$$

in this specific situation. It clearly holds

$$\mathbb{P}(A \cap B) = 1 - \mathbb{P}(A^{\complement} \cup B^{\complement}).$$

Using the union bound estimate we get

$$\mathbb{P}(A^{\complement} \cup B^{\complement}) \leq \mathbb{P}(A^{\complement}) + \mathbb{P}(B^{\complement}).$$

Theorem 2.3 implies

$$\mathbb{P}(B^{\complement}) \leq n^{1-r}.$$

And after noting

$$\mathbb{P}\left(A^{\complement}\right) = \mathbb{P}\left(\frac{1}{n}\left\|\mathbf{\Phi}_m^*\mathbf{\Phi}_m\right\|_{2\to 2} > F + \|\mathbf{\Lambda}\|_{2\to 2}\right) \leq \mathbb{P}\left(\left\|\frac{1}{n}\mathbf{\Phi}_m^*\mathbf{\Phi}_m - \mathbf{\Lambda}\right\|_{2\to 2} > F\right)$$

we infer from $\mathbf{\Phi}_m^*\mathbf{\Phi}_m = \sum_{i=1}^n \mathbf{y}^i \otimes \mathbf{y}^i$ and Proposition 3.8 that

$$\mathbb{P}\left(A^{\complement}\right) \leq \mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{y}^i \otimes \mathbf{y}^i - \mathbf{\Lambda}\right\|_{2\to 2} \geq F\right) \leq 2^{\frac{3}{4}}\, n^{1-r}.$$

In total we have

$$\mathbb{P}(A \cap B) \geq 1 - \eta n^{1-r}$$

with $\eta := 2^{3/4} + 1$. According to the proof of [11, Theorem 5.5] we need the function $T(\cdot, \mathbf{x}) := K(\cdot, \mathbf{x}) - \sum_{j=1}^{\infty} e_j(\cdot)\overline{e_j(\mathbf{x})}$, which denotes an element in $H(K)$. Its norm is given by

$$\|T(\cdot, \mathbf{x})\|_{H(K)}^2 := \langle T(\cdot, \mathbf{x}), T(\cdot, \mathbf{x})\rangle_{H(K)} = K(\mathbf{x}, \mathbf{x}) - \sum_{j=1}^{\infty} |e_j(\mathbf{x})|^2. \tag{5.2}$$

Note that the function in (5.2) is zero almost everywhere because of the fact that we have an equality sign in (4.6) due to our assumptions (separability of $H(K)$ and finite trace). Hence, $\mathbb{P}(C) = 1$ with

$$C := \left\{\mathbf{X} \in D^n \ : \ \sum_{k=1}^n \|T(\cdot, \mathbf{x}^k)\|_{H(K)} = 0\right\}.$$

Let now $\mathbf{X} \in A \cap B \cap C$. Then we can use for any $f \in H(K)$ with $\|f\|_{H(K)} \leq 1$ a similar argument as in [11, Theorem 5.5]

$$\begin{aligned}\|f - S^m_{\mathbf{X}} f\|^2_{L_2(D,\varrho_D)} &\leq \sigma_m^2 + \|(\mathbf{L}_m^*\mathbf{L}_m)^{-1}\mathbf{L}_m^*\|^2_{2\to 2} \cdot \sum_{k=1}^n \Big|\Big(f - P_{m-1}f\Big)(\mathbf{x}^k)\Big|^2 \\ &= \sigma_m^2 + \frac{2}{n}\|\mathbf{\Phi}_m\|^2_{2\to 2} + \frac{6\|K\|_\infty}{n}\sum_{k=1}^n \|T(\cdot, \mathbf{x}^k)\|_{H(K)} \\ &= \sigma_m^2 + \frac{2}{n}\|\mathbf{\Phi}_m\|_{2\to 2} \\ &\leq 2F + 3\sigma_m^2.\end{aligned}$$

This yields

$$\mathbb{P}\left(\sup_{\|f\|_{H(K)}\leq 1} \|f - S^m_{\mathbf{X}} f\|^2_{L_2(D,\varrho_D)} \leq 2F + 3\sigma_m^2\right) \geq 1 - \eta n^{1-r}$$

and therefore

$$\mathbb{P}\left(\sup_{\|f\|_{H(K)}\geq 1} \left\|f - S^m_{\mathbf{X}} f\right\|^2_{L_2(D,\varrho_D)} \leq 5\max\left\{\sigma_m^2, \frac{8r\log n}{n}T(m)\kappa^2\right\}\right) \geq 1 - \eta n^{1-r}.$$

□

In the sequel we consider a more general situation. The measure where the points are sampled will be adapted to the spectral properties of the embedding. This allows to specify the bound above in terms of the singular numbers of the embedding and benefit from their decay. Let us recall the density function from [13] which we will adapt to our framework as follows

$$\varrho_m(\mathbf{x}) = \frac{1}{2}\left(\frac{1}{m-1}\sum_{j=1}^{m-1}|\eta_j(\mathbf{x})|^2 + \frac{K(\mathbf{x},\mathbf{x}) - \sum_{j=1}^{m-1}|e_j(\mathbf{x})|^2}{\int_D K(\mathbf{x},\mathbf{x})d\varrho_D(\mathbf{x}) - \sum_{j=1}^{m-1}\lambda_j}\right). \qquad (5.3)$$

Algorithm 1 Weighted least squares approximation [5],[13],[11].

Input:	$\mathbf{X} = (\mathbf{x}^1, ..., \mathbf{x}^n) \in D^n$	set of distinct sampling nodes,
	$\mathbf{f} = (f(\mathbf{x}^1), ..., f(\mathbf{x}^n))^\top$	samples of f evaluated at the nodes from $\mathbf{X}$,
	$m \in \mathbb{N}$	$m < n$ such that the matrix $\tilde{\mathbf{L}}_m$ in (5.4) has full (column) rank.

Compute weighted samples $\boldsymbol{g} := (g_j)_{j=1}^n$ with $g_j := \begin{cases} 0, & \varrho_m(\mathbf{x}^j) = 0, \\ f(\mathbf{x}^j)/\sqrt{\varrho_m(\mathbf{x}^j)}, & \varrho_m(\mathbf{x}^j) \neq 0\,. \end{cases}$

Solve the over-determined linear system

$$\widetilde{\mathbf{L}}_m \cdot (\tilde{c}_1, ..., \tilde{c}_{m-1})^\top = \mathbf{g}\,,\ \widetilde{\mathbf{L}}_m := \Big(l_{j,k}\Big)_{j=1,k=1}^{n,m-1},\ l_{j,k} := \begin{cases} 0, & \varrho_m(\mathbf{x}^j) = 0, \\ \eta_k(\mathbf{x}^j)/\sqrt{\varrho_m(\mathbf{x}^j)}, & \varrho_m(\mathbf{x}^j) \neq 0, \end{cases} \tag{5.4}$$

via least squares (e.g. directly or via the LSQR algorithm [21]), i.e., compute

$$(\tilde{c}_1, ..., \tilde{c}_{m-1})^\top := (\widetilde{\mathbf{L}}_m^* \widetilde{\mathbf{L}}_m)^{-1}\, \widetilde{\mathbf{L}}_m^* \cdot \mathbf{g}.$$

Output: $\tilde{\mathbf{c}} = (\tilde{c}_1, ..., \tilde{c}_{m-1})^\top \in \mathbb{C}^{m-1}$ coefficients of the approximant $\widetilde{S}_{\mathbf{X}}^m f := \sum_{k=1}^{m-1} \tilde{c}_k \eta_k$.

We know from (4.6) that the sequence of singular numbers is square summable. We use the modified density function $\varrho_m(\cdot) : D \to \mathbb{R}$ which has been introduced in [11] as a version of the one from [13]. As above, the family $(e_j(\cdot))_{j\in\mathbb{N}}$ represents the eigenvectors of the non-vanishing eigenvalues of the compact self-adjoint operator $W_{\varrho_D} := \mathrm{Id}^* \circ \mathrm{Id} : H(K) \to H(K)$, the sequence $(\lambda_j)_{j\in\mathbb{N}}$ represents the ordered eigenvalues and finally $\eta_j := \lambda_j^{-1/2} e_j$.

Clearly, as a consequence of (4.5) the function ϱ_m is positive and defined pointwise for any $\mathbf{x} \in D$. Moreover, it can be computed precisely from the knowledge of $K(\mathbf{x}, \mathbf{x})$ and the first $m - 1$ eigenvalues and corresponding eigenvectors. It clearly holds that $\int_D \varrho_m(\mathbf{x}) d\varrho_D(\mathbf{x}) = 1$. Here is one of the main results of this paper (note that Theorem 1.3 from the introduction is a simple reformulation of the theorem below).

Theorem 5.2 *Let $H(K)$ be a separable reproducing kernel Hilbert space of complex-valued functions defined on D such that*

$$\int_D K(\mathbf{x}, \mathbf{x}) d\varrho_D(\mathbf{x}) < \infty$$

for some measure ϱ_D on D, where $(\sigma_k)_k$ denotes the non-increasing sequence of singular numbers of the embedding $\mathrm{Id} : H(K) \to L_2(D, \varrho_D)$. Then we have for $n \in \mathbb{N}$ and

$$m := \left\lfloor \frac{n}{14r \log n} \right\rfloor \tag{5.5}$$

the bound

$$\mathbb{P}\left(\sup_{\|f\|_{H(K)} \leq 1} \|f - \widetilde{S}_{\mathbf{X}}^m f\|_{L_2(D,\varrho_D)}^2 \leq 5 \max\left\{ \sigma_m^2, \frac{16 r \kappa^2 \log n}{n} \sum_{j=m}^{\infty} \sigma_j^2 \right\} \right) \geq 1 - \eta n^{1-r},$$

where $\mathbf{X}$ *is sampled i.i.d. according to* $\varrho_m(\mathbf{x})d\varrho_D(\mathbf{x})$ *in* (5.3) *above and* $\eta = 2^{\frac{3}{4}} + 1$, $\kappa = \frac{1+\sqrt{5}}{2}$.

Proof This result is a consequence of Theorem 5.1 above which is applied to the newly constructed probability measure $\varrho_m(\cdot)$ together with

$$\widetilde{K}_m(\mathbf{x}, \mathbf{y}) := \frac{K(\mathbf{x}, \mathbf{y})}{\sqrt{\varrho_m(\mathbf{x})}\sqrt{\varrho_m(\mathbf{y})}}. \tag{5.6}$$

We observe

$$\sup_{\|f\|_{H(K)}\le 1} \big\| f - \widetilde{S}^m_{\mathbf{X}} f \big\|_{L_2(D,\varrho_D)} \le \sup_{\|g\|_{H(\widetilde{K}_m)}\le 1} \big\| g - S^m_{\mathbf{X}} g \big\|_{L_2(D,\mu_m)},$$

$\widetilde{N}(m) \le 2(m-1)$, and $\widetilde{T}(m) \le 2\sum_{j=m}^{\infty} \sigma_j^2$, see [11, Thms. 5.5, 5.8]. Applying Theorem 5.1 leads to the stated bound. □

6 Sampling discretization of the L_2-norm

Motivated from supervised learning theory, see e.g. [6], one is interested in uniform bounds for the following version of the "defect function"

$$L_{\mathbf{X}}(f) := \int_D |f(\mathbf{x})|^2\, d\varrho_D(\mathbf{x}) - \frac{1}{n}\sum_{j=1}^{n} |f(\mathbf{x}^j)|^2$$

with respect to f belonging to some hypothesis space $\mathcal{H}$ which is usually embedded into $\mathcal{C}(D)$, the continuous functions on the domain D. From a more classical perspective, authors were interested in discretizing L_p-norms using Marcinkiewicz–Zygmund inequalities. This subject has been recently studied systematically by Temlyakov [25], see also Gröchenig [9]. The following theorem will be an immediate implication of our concentration result in Theorem 1.1.

Theorem 6.1 *Let* ϱ_D *denote a probability measure on the measurable subset* $D \subset \mathbb{R}^d$ *and* $H(K)$ *be a separable reproducing kernel Hilbert space with kernel* $K(\cdot,\cdot)$ *such that*

$$\|K\|_\infty := \sup_{\mathbf{x}\in D} \sqrt{K(\mathbf{x},\mathbf{x})} < \infty$$

(*equivalently*, *the unit ball in* $H(K)$ *is uniformly bounded in* ℓ_∞). *Then we have for* $0 < t \le 1$

$$\mathbb{P}\left(\sup_{\|f\|_{H(K)}\le 1} \left| \int_D |f(\mathbf{x})|^2\, d\varrho_D(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^{n} |f(\mathbf{x}^i)|^2 \right| > t\|\mathrm{Id}\|_{K,2}^2\right) \le 2^{3/4} n \exp\left(-\frac{nt^2\|\mathrm{Id}\|_{K,2}^2}{21\|K\|_\infty^2}\right).$$

If we fix $r > 1$ the above bound can be reformulated as

$$\sup_{\|f\|_{H(K)}\le 1}\left|\int_D |f(\mathbf{x})|^2\, d\varrho_D(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^n |f(\mathbf{x}^i)|^2\right| \le \|\mathrm{Id}\|_{H(K)\to L_2}\|K\|_\infty\sqrt{\frac{21r\log n}{n}}$$

with probability exceeding $1-2n^{1-r}$ and n sufficiently large, namely

$$\frac{n}{\log n} \ge \frac{21r\|K\|_\infty^2}{\|\mathrm{Id}\|_{K,2}^2}.$$

The nodes $\mathbf{X} = (\mathbf{x}^1,\ldots,\mathbf{x}^n)$ are randomly drawn according to the product measure $(d\varrho_D(\mathbf{x}))^n$.

Proof Fix f with $\|f\|_{H(K)} \le 1$ and put $M := \|K\|_\infty$. Due to the L_2-identity

$$f = \sum_{i=1}^\infty \sigma_i\langle f, e_i\rangle\eta_i$$

we find

$$\int_D |f(\mathbf{x})|^2\, d\varrho_D(\mathbf{x}) = \sum_{i=1}^\infty \sigma_i^2|\langle f, e_i\rangle|^2 = \langle\hat{\mathbf{f}}, \boldsymbol{\Lambda}\hat{\mathbf{f}}\rangle_{\ell_2}$$

with $\hat{\mathbf{f}} = (\langle f, e_k\rangle_{H(K)})_{k\in\mathbb{N}}$ and $\boldsymbol{\Lambda} = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots)$. Note that $\|\boldsymbol{\Lambda}\|_{2\to 2} = \|\mathrm{Id}\|_{K,2}^2$. Furthermore, putting

$$\mathbf{y} = (e_1(\mathbf{x}^i), e_2(\mathbf{x}^i), \ldots, e_k(\mathbf{x}^i), \ldots)^\top, \quad i = 1,\ldots,n,$$

we find

$$\frac{1}{n}\sum_{i=1}^n |f(\mathbf{x}^i)|^2 = \left\langle\hat{\mathbf{f}}, \left(\frac{1}{n}\sum_{i=1}^n \mathbf{y}^i\otimes\mathbf{y}^i\right)\hat{\mathbf{f}}\right\rangle_{\ell_2}$$

holds almost surely since $\mathrm{tr}_0(K) = 0$ due to the separability of $H(K)$, see the paragraph after (4.6). In fact, the identity fails on a nullset $A \subset D^n$, which is independent of f. This follows by using the same arguments as after (5.2). Hence,

$$\begin{aligned}
&\sup_{\|f\|_{H(K)}\le 1}\left|\int_D |f(\mathbf{x})|^2\, d\varrho_D(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^n |f(\mathbf{x}^i)|^2\right| \\
&= \sup_{\|\hat{\mathbf{f}}\|_{\ell_2}\le 1}\left|\langle \hat{\mathbf{f}}, \mathbf{\Lambda}\hat{\mathbf{f}}\rangle_{\ell_2} - \left\langle \hat{\mathbf{f}}, \left(\frac{1}{n}\sum_{i=1}^n \mathbf{y}^i\otimes\mathbf{y}^i\right)\hat{\mathbf{f}}\right\rangle_{\ell_2}\right| \\
&= \sup_{\|\hat{\mathbf{f}}\|_{\ell_2}\le 1}\left|\left\langle \hat{\mathbf{f}}, \left(\mathbf{\Lambda} - \frac{1}{n}\sum_{i=1}^n \mathbf{y}^i\otimes\mathbf{y}^i\right)\hat{\mathbf{f}}\right\rangle_{\ell_2}\right| \\
&= \left\|\mathbf{\Lambda} - \frac{1}{n}\sum_{i=1}^n \mathbf{y}^i\otimes\mathbf{y}^i\right\|_{2\to 2} \\
&\le t\|\mathbf{\Lambda}\|_{2\to 2}
\end{aligned}$$

with probability exceeding $1-\exp(-t^2 n/(21\tilde{M}^2))$ by Theorem 1.1. Here $\tilde{M}^2 = M^2/\|\mathbf{\Lambda}\|_{2\to 2}$. Hence, we may choose $t=\sqrt{21\tilde{M}^2 r\log(n)/n}\le 1$ to finally get

$$\sup_{\|f\|_{H(K)}\le 1}\left|\int_D |f(\mathbf{x})|^2\, d\varrho_D(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^n |f(\mathbf{x}^i)|^2\right| \le \sqrt{21\|\mathbf{\Lambda}\|_{2\to 2}M^2 r\frac{\log n}{n}}. \quad (6.1)$$

□

Remark 6.2 (i) The uniform boundedness (in ℓ_∞) of a function class is sometimes also called M-boundedness in the learning theory literature. It represents a common assumption there to analyze the defect function. In fact, uniform bounds on the defect function are proved by using the concept of covering numbers of the unit ball of $H(K)$ in $\ell_\infty(D)$, see [12,25]. In the above theorem covering number estimates were not used at all.

(ii) The quantity $\|K\|_\infty$ may be replaced by $\|K\|_{L_\infty(D,\varrho_D)}$, i.e., the essential supremum with respect to the probability measure ϱ_D. Since $\|K\|_{L_\infty(D,\varrho_D)}$ might by smaller than $\|K\|_\infty$ we obtain a slight improvement.

Now we want to get rid of the uniform boundedness of the class and only assume the finite trace. We have to change the sampling measure and modify the norm discretization operator by incorporating weights. The corresponding theorem reads as follows.

Theorem 6.3 *Let ϱ_D denote an arbitrary measure on the measurable subset $D\subset\mathbb{R}^d$ and $H(K)$ be a separable reproducing kernel Hilbert space with (Hermitian) kernel $K(\cdot,\cdot)$ such that*

$$\operatorname{tr}(K) := \int_D K(\mathbf{x},\mathbf{x})\, d\varrho_D(\mathbf{x}) < \infty.$$

We define the probability density function

$$\nu(\mathbf{x}) := \frac{K(\mathbf{x},\mathbf{x})}{\operatorname{tr}(K)}$$

and sample $\mathbf{X} = (\mathbf{x}^1,\dots,\mathbf{x}^n)$ *from the product measure* $(\nu(\mathbf{x}))d\varrho_D(\mathbf{x})^n$. *Then we have*

$$\mathbb{P}\left(\sup_{\|f\|_{H(K)}\le 1}\left|\int_D |f(\mathbf{x})|^2\, d\varrho_D(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^n \frac{|f(\mathbf{x}^i)|^2}{\nu(\mathbf{x}^i)}\right| > t\|\mathrm{Id}\|_{K,2}^2\right) \le 2^{3/4} n \exp\left(-\frac{nt^2\|\mathrm{Id}\|_{K,2}^2}{21\operatorname{tr}(K)}\right).$$

If we fix $r > 1$ *then this result can be reformulated as*

$$\sup_{\|f\|_{H(K)}\le 1}\left|\int_D |f(\mathbf{x})|^2\, d\varrho_D(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^n \frac{|f(\mathbf{x}^i)|^2}{\nu(\mathbf{x}^i)}\right| \le \sqrt{21\operatorname{tr}(K)\|\mathrm{Id}\|^2 r\frac{\log n}{n}}$$

with probability exceeding $1 - 2n^{1-r}$ *and n sufficiently large*, *namely*

$$\frac{n}{\log n} \ge \frac{21r\operatorname{tr}(K)}{\|\mathrm{Id}\|_{K,2}^2}.$$

Proof We want to apply Theorem 6.1. Let us define the normalized kernel

$$\tilde{K}(\mathbf{x},\mathbf{y}) := \frac{K(\mathbf{x},\mathbf{y})}{\sqrt{\nu(\mathbf{x})}\sqrt{\nu(\mathbf{y})}}.$$

Then $\|\tilde{K}\|_\infty = \mathrm{tr(K)}$ and

$$\begin{aligned}&\sup_{\|f\|_{H(K)}\le 1}\left|\int_D |f(\mathbf{x})|^2\, d\varrho_D(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^n \frac{|f(\mathbf{x}^i)|^2}{\nu(\mathbf{x}^i)}\right| \\ &= \sup_{\|f\|_{H(\tilde{K})}\le 1}\left|\int_D |f(\mathbf{x})|^2\, \nu(\mathbf{x})d\varrho_D(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^n |f(\mathbf{x}^i)|^2\right|.\end{aligned} \tag{6.2}$$

It remains to note that

$$\|\mathrm{Id}\|_{K,2,d\varrho_D} = \|\mathrm{Id}\|_{\tilde{K},2,\nu(\cdot)d\varrho_D(\cdot)}$$

and we may apply Theorem 6.1. □

7 Non-separable RKHS

Now we deal with a more general situation and drop the separability assumption for $H(K)$. We only assume the finite trace property (1.5). For this purpose we define the new density function

$$\varrho_m(\mathbf{x}) = \frac{1}{3}\left(\frac{\sum_{j=1}^{m-1}|\eta_j(\mathbf{x})|^2}{m-1} + \frac{\sum_{j=m}^{\infty}|e_j(\mathbf{x})|^2}{\sum_{j=m}^{\infty}\lambda_j} + \frac{K_0(\mathbf{x},\mathbf{x})}{\operatorname{tr}(K^0)}\right) \tag{7.1}$$

and the operator $\widetilde{S}_{\mathbf{X}}^m$ from Algorithm 1. Clearly, it holds $\int \varrho_m(\mathbf{x})d\varrho_D(\mathbf{x}) = 1$. Theorem 1.2 provides the bound

$$\sup_{\|f\|_{H(K)}\le 1} \|f - \widetilde{S}_{\mathbf{X}}^m f\|_{L_2(D,\varrho_D)}^2 \le C \max\left\{\sigma_m^2, \frac{r\log n}{n}\sum_{j=m}^{\infty}\sigma_j^2, \frac{\operatorname{tr}(K^0)}{n}\right\} \tag{7.2}$$

with an absolute constant $C > 0$ and $m := \lfloor n/(14r\log n)\rfloor$. Note that this result improves on a result in Wasilkowski and Woźniakowski [31], see also Novak and Woźniakowski [19, Thm. 26.10]. The authors in [31, Thm. 26.10] constructed a recovery operator using n samples having squared worst-case error not greater than

$$\min\left\{\sigma_\ell^2 + \frac{\operatorname{tr}(K)\ell}{n} \ : \ \ell = 1, 2, 3, \dots\right\}.$$

If we for instance assume that $\sum_{k=1}^{\infty}\sigma_k^p < \infty$ for some $0 < p \le 2$ then we may balance $\ell \asymp n^{p/(p+2)}$ to obtain a rate of $\mathcal{O}(n^{-1/(p+2)})$. In Theorem 1.2, see (7.2), we obtain a rate of $o(\sqrt{(\log n)/n})$ already for $p = 2$. In case $p < 2$ we obtain $\mathcal{O}(n^{-1/2})$. It seems that, in general, the decay properties of the singular values have a rather weak influence on the recovery bound compared to the separable case, where it is much better than $\mathcal{O}(n^{-1/2})$. A lower bound showing that we can not essentially improve on $\mathcal{O}(n^{-1/2})$ with the above algorithm will be provided in [17].

Proof of Theorem 1.2

Proof *Step 1.* Let us assume that $M := \|K\|_\infty = \sup_{\mathbf{x}\in X}\sqrt{K(\mathbf{x},\mathbf{x})} < \infty$. By the spectral theorem we can decompose

$$H(K) = N(\mathrm{Id}) \oplus \overline{\operatorname{span}\{e_1(\cdot), e_2(\cdot), \dots\}}$$

where N is the nullspace of the embedding. Let us now define

$$K^1(\mathbf{x},\mathbf{y}) = \sum_{j=1}^{\infty} e_j(\mathbf{x})\overline{e_j(\mathbf{y})}$$

and

$$K^0(\mathbf{x},\mathbf{y}) = K(\mathbf{x},\mathbf{y}) - K^1(\mathbf{x},\mathbf{y}).$$

Therefore, $K^0(\mathbf{x}, \mathbf{y})$ is the reproducing kernel of the nullspace $N(\mathrm{Id})$ and $\sup_{\mathbf{x}\in X} \sqrt{K^0(\mathbf{x}, \mathbf{x})} =: M_0 \le M < \infty$. We estimate

$$\begin{aligned}
&\sup_{\|f\|_{H(K)}\le 1} \left\| f - S^m_{\mathbf{X}} f \right\|_{L_2(D,\varrho_D)} \\
&\quad\le \sup_{\|g\|_{H(K^0)}\le 1} \left\| g - S^m_{\mathbf{X}} g \right\|_{L_2(D,\varrho_D)} + \sup_{\|g\|_{H(K^1)}\le 1} \left\| g - S^m_{\mathbf{X}} g \right\|_{L_2(D,\varrho_D)}.
\end{aligned} \tag{7.3}$$

The second summand can be treated with Theorem 5.1. The space $H(K^1)$ is separable and $K^1(\mathbf{x}, \mathbf{x})$ is bounded. Theorem 5.1 gives

$$\sup_{\|g\|_{H(K^1)}\le 1} \left\| g - S^m_{\mathbf{X}} g \right\|^2_{L_2(D,\varrho_D)} \le 5 \max\left\{ \sigma_m^2, \frac{8r \log n}{n} T(m)\kappa^2 \right\} \tag{7.4}$$

with probability at least $1 - \eta n^{1-r}$ whenever (5.1) holds. The number κ is the same as in Proposition 3.8.

Note that all the functions in $H(K^0)$ are zero in $L_2(D, \varrho_D)$ since this space is the nullspace of the embedding. Hence

$$\begin{aligned}
\sup_{\|g\|_{H(K^0)}\le 1} \left\| g - S^m_{\mathbf{X}} g \right\|^2_{L_2(D,\varrho_D)} &= \sup_{\|g\|_{H(K^0)}\le 1} \left\| S^m_{\mathbf{X}} g \right\|^2_{L_2(D,\varrho_D)} \\
&\le \frac{2}{n} \sup_{\|g\|_{H(K^0)}\le 1} \sum_{i=1}^n |g(\mathbf{x}^i)|^2
\end{aligned}$$

holds for the same $\mathbf{X} = (\mathbf{x}^1, \dots, \mathbf{x}^n)$ for which (7.4) holds. We only need the operator norm of $\mathbf{H}_m$ (see (2.3)) to be larger than $\frac{1}{2}$ which comes from Theorem 2.1. At this point we employ the "Representer Theorem" from learning theory, see for instance [4, Theorem 5.5]. We claim that

$$\sup_{\|g\|_{H(K^0)}\le 1} \sum_{i=1}^n |g(\mathbf{x}^i)|^2 = \sup_{\boldsymbol{\alpha}^\top K^0[\mathbf{X}]\overline{\boldsymbol{\alpha}}\le 1} \sum_{i=1}^n |g(\mathbf{x}^i)|^2, \tag{7.5}$$

where $\left(K^0(\mathbf{x}^i, \mathbf{x}^j)\right)_{i,j=1}^n$ is the kernel taken at the points from $\mathbf{X}$ and $g(\mathbf{x}) = \sum_{i=1}^n \alpha_i K^0(\mathbf{x}, \mathbf{x}^i)$

In other words, we can reduce the problem to the finite dimensional space $\operatorname{span}\left\{K^0(\cdot, \mathbf{x}^1), \dots, K^0(\cdot, \mathbf{x}^n)\right\}$. Note that

$$\begin{aligned}
\|g\|^2_{H(K^0)} &= \sum_{i=1}^n \sum_{j=1}^n K^0(\mathbf{x}^i, \mathbf{x}^j)\alpha_i\overline{\alpha_j} \\
&= \boldsymbol{\alpha}^\top K^0[\mathbf{X}]\overline{\boldsymbol{\alpha}} \\
&\le 1.
\end{aligned}$$

The reason is that $g \in H(K^0)$ can be decomposed into $g = g_1 + g_2$ with $g_1 \perp g_2$ and $g_1 = Pg$, the orthogonal projection onto $\operatorname{span}\big\{K^0(\cdot, \mathbf{x}^1), \ldots, K^0(\cdot, \mathbf{x}^n)\big\}$. Due to $g_1 \perp g_2$, we have that $\langle g_2, K^0(\cdot, \mathbf{x}^i)\rangle = 0 = g_2(\mathbf{x}^i)$ for all i. Hence $\sum_{i=1}^n |g(\mathbf{x}^i)|^2 = \sum_{i=1}^n |g_1(\mathbf{x}^i)|^2$ and $\|g_1\|^2_{H(K^0)} \le 1$. Therefore

$$\begin{aligned}\sup_{\|g\|_{H(K^0)}\le 1} \big\|g - S_{\mathbf{X}}^m g\big\|^2_{L_2(D,\varrho_D)} &\le \frac{2}{n} \sup_{\|g\|_{H(K^0)}\le 1} \sum_{i=1}^n |g(\mathbf{x}^i)|^2 \\ &\le \frac{2}{n} \sup_{\boldsymbol{\alpha}^\top K^0[\mathbf{X}]\overline{\boldsymbol{\alpha}}\le 1} \sum_{i=1}^n \left|\sum_{j=1}^n \alpha_j K^0(\mathbf{x}^i, \mathbf{x}^j)\right|^2. \qquad (7.6)\end{aligned}$$

Since $H(K^0)$ is the nullspace of the embedding we know that $K^0(\mathbf{x}^i, \mathbf{x}^j)$ is zero almost surely for $i \ne j$. We can therefore continue to estimate (7.6) by

$$\begin{aligned}\frac{2}{n} \sup_{\alpha^\top K[\mathbf{X}]\overline{\alpha}\le 1} \sum_{i=1}^n \left|\sum_{j=1}^n \alpha_j K^0(\mathbf{x}^i, \mathbf{x}^j)\right|^2 &= \frac{2}{n} \sup_{\boldsymbol{\alpha}^\top K[\mathbf{X}]\overline{\boldsymbol{\alpha}}\le 1} \sum_{i=1}^n |\alpha_i|^2 \big|K^0(\mathbf{x}^i, \mathbf{x}^i)\big|^2 \\ &\le \frac{2M_0^2}{n} \sup_{\boldsymbol{\alpha}^\top K[\mathbf{X}]\overline{\boldsymbol{\alpha}}\le 1} \sum_{i=1}^n |\alpha_i|^2 \big|K^0(\mathbf{x}^i, \mathbf{x}^i)\big| \\ &\le \frac{2M_0^2}{n}\end{aligned} \qquad (7.7)$$

since we have almost surely

$$\begin{aligned}\sup_{\boldsymbol{\alpha}^\top K^0[\mathbf{X}]\overline{\boldsymbol{\alpha}}\le 1} \sum_{i=1}^n |\alpha_i|^2 \big|K^0(\mathbf{x}^i, \mathbf{x}^i)\big| &= \sum_{i=1}^n \sum_{j=1}^n \alpha_i \overline{\alpha_j} K^0(\mathbf{x}^i, \mathbf{x}^j) \\ &= \boldsymbol{\alpha}^\top K^0[\mathbf{X}]\overline{\boldsymbol{\alpha}} \\ &\le 1.\end{aligned}$$

This leads to

$$\begin{aligned}&\sup_{\|f\|_{H(K)}\le 1} \big\|f - S_{\mathbf{X}}^m f\big\|^2_{L_2(D,\varrho_D)} \\ &\le \left(\sup_{\|g\|_{H(K^0)}\le 1} \big\|g - S_{\mathbf{X}}^m g\big\|_{L_2(D,\varrho_D)} + \sup_{\|g\|_{H(K_m^1)}\le 1} \big\|g - S_{\mathbf{X}}^m g\big\|_{L_2(D,\varrho_D)}\right)^2 \\ &\le \left(\sqrt{\frac{1}{4\kappa^2}} + \sqrt{5}\right)^2 \max\left\{\sigma_m^2, \frac{8r\log n}{n} T(m)\kappa^2, \frac{8M_0^2\kappa^2}{n}\right\} \\ &\le 7 \max\left\{\sigma_m^2, \frac{8r\log n}{n} T(m)\kappa^2, \frac{8M_0^2\kappa^2}{n}\right\} \qquad (7.8)\end{aligned}$$

with probability exceeding $1 - \eta n^{1-r}$.

Step 2. We define the measure $d\mu_m(\mathbf{x}) = \varrho_m(\mathbf{x})d\varrho_D(\mathbf{x})$ as well as the kernel $\widetilde{K}_m(\mathbf{x},\mathbf{y})$ as in (5.6) and $\widetilde{K}_m^0$, $\widetilde{K}_m^1$ accordingly. This gives

$$\sup_{\|f\|_{H(K)}\le 1} \|f - \widetilde{S}_{\mathbf{X}}^m f\|_{L_2(D,\varrho_D)} \le \sup_{\|g\|_{H(\widetilde{K}_m)}\le 1} \|g - \widetilde{S}_{\mathbf{X}}^m g\|_{L_2(D,\mu_m)}. \tag{7.9}$$

We apply the results from Step 1 to the right-hand side. Hence, we have to know the bound for $\widetilde{K}_m^0$, $\widetilde{N}(m)$ and $\widetilde{T}(m)$, where the latter quantities are associated to $\widetilde{K}_m^1$. We will now show that $\widetilde{K}_m^0$ can be bounded by $3\,\mathrm{tr}(K^0)$. In fact,

$$\begin{aligned}\widetilde{K}_m^0(\mathbf{x},\mathbf{x}) &= \frac{K_m^0(\mathbf{x},\mathbf{x})}{\varrho_m(\mathbf{x})}\\ &= \frac{K(\mathbf{x},\mathbf{x}) - \sum_{k=1}^{\infty}|e_k(\mathbf{x})|^2}{\frac{1}{3}\left(\frac{\sum_{j=1}^{m-1}|\eta_j(\mathbf{x})|^2}{m-1} + \frac{\sum_{j=m}^{\infty}|e_j(\mathbf{x})|^2}{\sum_{j=m}^{\infty}\lambda_j} + \frac{K^0(\mathbf{x},\mathbf{x})}{\mathrm{tr}(K^0)}\right)}\\ &\le 3\ \mathrm{tr}(K^0).\end{aligned} \tag{7.10}$$

Hence, we have $\widetilde{M}_0^2 = 3\,\mathrm{tr}(K^0)$. By the same arguments as in the proof of Theorem 5.2, see also [11, Thm. 5.7], we get $\widetilde{N}(m) \le 3(m-1)$ and $\widetilde{T}(m) \le 3\sum_{j=m}^{\infty}\sigma_j^2$. Plugging this into (7.8) gives

$$\sup_{\|g\|_{H(\widetilde{K}_m)}\le 1} \|g - \widetilde{S}_{\mathbf{X}}^m g\|_{L_2(D,\mu_m)}^2 \le 7\max\left\{\sigma_m^2, \frac{24\kappa^2 r\log n}{n}\sum_{j=m}^{\infty}\sigma_j^2, \frac{24\kappa^2\,\mathrm{tr}(K^0)}{n}\right\}$$

$$\sup_{\|f\|_{H(K)}\le 1} \|f - \widetilde{S}_{\mathbf{X}}^m f\|_{L_2(D,\varrho_D)}^2 \le 441\max\left\{\sigma_m^2, \frac{r\log n}{n}\sum_{j=m}^{\infty}\sigma_j^2, \frac{\mathrm{tr}(K^0)}{n}\right\}.$$

□

What concerns a counterpart of the L_2-norm discretization result for general RKHS having finite trace we can prove the following.

Theorem 7.1 *Let $H(K)$ be a RKHS and ϱ_D be a measure on $D \subset \mathbb{R}^d$. If*

(i) *If $\|K\|_\infty := \sup_{\mathbf{x}\in D}\sqrt{K(\mathbf{x},\mathbf{x})} < \infty$ and ϱ_D denotes a probability measure on D, where $\mathbf{x}^i$, $i = 1,\ldots,n$, are drawn i.i.d. according to ϱ_D then*

$$\sup_{\|f\|_{H(K)}\le 1}\left|\int_D |f(\mathbf{x})|^2 d\varrho_D(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^{n}|f(\mathbf{x}^i)|^2\right| \le 8\sqrt{r\frac{\log(n)}{n}}\|K\|_\infty^2$$

holds with probability at least $1-2n^{1-r}$ if n is large enough, i.e. $\frac{n}{\log n} \ge \frac{21r\|K\|_\infty^2}{\|\mathrm{Id}\|_{K,2}^2}$.

(ii) *If* $\operatorname{tr}(K) = \int_D K(\mathbf{x},\mathbf{x}) d\varrho_D(\mathbf{x}) < \infty$ *then*

$$\sup_{\|f\|_{H(K)}\le 1} \left| \int_D |f(\mathbf{x})|^2 d\varrho_D(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^n \frac{|f(\mathbf{x}^i)|^2}{\nu(\mathbf{x}^i)} \right| \le 8\operatorname{tr}(K)\sqrt{r\frac{\log(n)}{n}}$$

holds with probability at least $1-2n^{1-r}$ *and* $\frac{n}{\log n} \ge \frac{21 r \operatorname{tr}(K)}{\|\mathrm{Id}\|^2_{K;2}}$ *where* $\nu(\mathbf{x}) = \frac{K(\mathbf{x},\mathbf{x})}{\operatorname{tr}(K)}$ *and* $\mathbf{x}^i$, $i = 1,\dots,n$, *are drawn i.i.d. according to* $\nu(\mathbf{x})d\varrho_D(\mathbf{x})$.

Proof Since we may have that $\operatorname{tr}_0(K) > 0$ the decomposition $K(\mathbf{x},\mathbf{y}) = K^0(\mathbf{x},\mathbf{y}) + K^1(\mathbf{x},\mathbf{y})$ leads to a "non-trivial" Kernel $K^0(\mathbf{x},\mathbf{y})$. We estimate in case (i):

$$\begin{aligned}
&\sup_{\|f\|_{H(K)}\le 1} \left| \int_D |f(\mathbf{x})|^2 d\varrho_D(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^n |f(\mathbf{x}^i)|^2 \right| \\
&= \sup_{\|f\|_{H(K)}\le 1} \left| \int_D |f_0(\mathbf{x}) + f_1(\mathbf{x})|^2 d\varrho_D(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^n |f_0(\mathbf{x}^i) + f_1(\mathbf{x}^i)|^2 \right| \\
&= \sup_{\|f\|_{H(K)}\le 1} \left| \int_D |f_1(\mathbf{x})|^2 d\varrho_D(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^n |f_0(\mathbf{x}^i)|^2 - \frac{2}{n}\sum_{i=1}^n |f_0(\mathbf{x}^i) f_1(\mathbf{x}^i)| - \frac{1}{n}\sum_{i=1}^n |f_1(\mathbf{x}^i)|^2 \right| \\
&\le \sup_{\|f\|_{H(K^1)}\le 1} \left| \int_D |f_1(\mathbf{x})|^2 d\varrho_D(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^n |f_1(\mathbf{x}^i)|^2 \right| &(7.11)\\
&\quad + \sup_{\|f\|_{H(K^0)}\le 1} \frac{1}{n}\sum_{i=1}^n |f_0(\mathbf{x}^i)|^2 &(7.12)\\
&\quad + \sup_{\|f\|_{H(K)}\le 1} \frac{2}{n}\left(\sum_{i=1}^n |f_0(\mathbf{x}^i)|^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^n |f_1(\mathbf{x}^i)|^2\right)^{\frac{1}{2}} &(7.13)
\end{aligned}$$

and note that (7.11) $\le \sqrt{21\|\mathrm{Id}\|^2 M^2 r \frac{\log n}{n}}$ by Theorem 6.1 with probability at least $1-2n^{1-r}$, where $M := \|K\|_\infty$. To estimate (7.12) we use the same reasoning leading to (7.5) and get

$$\sup_{\|f\|_{H(K^0)}\le 1} \frac{1}{n}\sum_{i=1}^n |f_0(\mathbf{x}^i)|^2 \le \frac{1}{n}M^2, \tag{7.14}$$

where we used $K(\mathbf{x},\mathbf{x}) \le M^2$. We also use (7.14) in order to estimate (7.13). It holds

$$\begin{aligned}
\sup_{\|f\|_{H(K)}\le 1} \frac{2}{n}\left(\sum_{i=1}^n |f_0(\mathbf{x}^i)|^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^n |f_1(\mathbf{x}^i)|^2\right)^{\frac{1}{2}} &\le \frac{2}{\sqrt{n}} M \left(\frac{1}{n}\sum_{i=1}^n |f_1(\mathbf{x}^i)|^2\right)^{\frac{1}{2}} \\
&\le \frac{2M^2}{\sqrt{n}}.
\end{aligned}$$

In total we estimate

$$(7.11) + (7.12) + (7.13) \leq \sqrt{21r}M^2\sqrt{\frac{\log(n)}{n}} + \frac{M^2}{n} + \frac{2M^2}{\sqrt{n}}$$
$$\leq 8M^2\sqrt{\frac{r\log(n)}{n}}.$$

To prove (ii) we use the same technique as in Theorem 6.3 replacing $\frac{1}{n}\sum_{i=1}^{n}|f(\mathbf{x}^i)|^2$ with $\frac{1}{n}\sum_{i=1}^{n}\frac{|f(\mathbf{x}^i)|^2}{\nu(\mathbf{x}^i)}$ where $\nu(\mathbf{x}) = \frac{K(\mathbf{x},\mathbf{x})}{\operatorname{tr}(K)}$ and also M^2 by $\operatorname{tr}(K)$ we can reduce everything to case (i). □

Acknowledgements T.U. would like to acknowledge support by the DFG Ul-403/2-1. The authors would like to thank T. Kühn, V.N. Temlyakov, M. Ullrich, K. Pozharska and two anonymous referees for comments and further references.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

1. Berlinet, A., Thomas-Agnan, C.: *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic Publishers, Boston, MA, (2004). With a preface by Persi Diaconis
2. Buchholz, A.: Operator Khintchine inequality in non-commutative probability. Ann. of Math. **319**, 1–16 (2001)
3. Buchholz, A.: Optimal constants in Khintchine type inequalities for Fermions, Rademachers and q-Gaussian operators. Bulletin Polish Acad. Sci. Math. **53**(3), 315–321 (2005)
4. Christmann, A., Steinwart, I.: Support Vector Machines. Springer, (2008)
5. Cohen, A., Migliorati, G.: Optimal weighted least-squares methods. SMAI J. Comput. Math. **3**, 181–203 (2017)
6. Cucker, F., Zhou, D.X.: Learning Theory. Cambridge University Press, An Approximation Theory Viewpoint (2007)
7. Dirksen, S.: *Noncommutative and vector-valued Rosenthal inequalities*. Dissertation, Delft Institute of Applied Mathematics, (2011)
8. Dũng, D., Temlyakov, V.N., Ullrich, T.: Hyperbolic Cross Approximation. Advanced Courses in Mathematics. CRM Barcelona, Birkhäuser/Springer (2019)
9. Gröchenig, K.: Sampling, Marcinkiewicz-Zygmund inequalities, approximation, and quadrature rules. J. Approx. Theory **257**,(2020)
10. Hein, M., Bousquet, O.: Kernels, associated structures and generalizations. Technical Report 127, Max Planck Institute for Biological Cybernetics, Tübingen, Germany, (2004)
11. Kämmerer, L., Ullrich, T., Volkmer, T.: Worst-case recovery guarantees for least squares approximation using random samples. Constr. Approx., to appear, arXiv:1911.10111v3
12. Konyagin, S.V., Temlyakov, V.N.: The entropy in learning theory. Error estimates. Constr. Approx. **25**(1), 1–27 (2007)

13. Krieg, D., Ullrich, M.: Function values are enough for L_2-approximation. *Found. Comp. Math.*, to appear, arXiv:math/1905.02516v3
14. Krieg, D., Ullrich, M.: Function values are enough for L_2-approximation. Part II. *J. Complexity*, to appear, arXiv:2011.01779
15. Ledoux, M., Talagrand, M.: Probability in Banach Spaces. Springer, (1991)
16. Mendelson, S., Pajor, A.: On singular values of matrices with independent rows. Bernoulli **12**, 761–773 (2006)
17. Moeller, M.: Norm-concentration results for infinite random matrices with independent rows. Bachelor's thesis, Faculty of Mathematics, TU Chemnitz (2021)
18. Nagel, N., Schäfer, M., Ullrich, T.: A new upper bound for sampling numbers. *Found. Comp. Math.*, to appear, arXiv:2010.00327
19. Novak, E., Woźniakowski, H.: *Tractability of multivariate problems. Volume III: Standard information for operators*, volume 18 of *EMS Tracts in Mathematics*. European Mathematical Society (EMS), Zürich, (2012)
20. Oliveira, R.I.: Sums of random Hermitian matrices and an inequality by Rudelson. Electr. Comm. Probab. **15**, 203–212 (2010)
21. Paige, C.C., Saunders, M.A.: LSQR: An algorithm for sparse linear equations and sparse least squares. ACM Trans. Math. Software **8**, 43–71 (1982)
22. Rauhut, H.: Compressive sensing and structured random matrices. In: Fornasier, M. (ed.) Theoretical Foundations and Numerical Methods for Sparse Recovery, volume 9 of Radon Series on Computational and Applied Mathematics. de Gruyter, Berlin (2010)
23. Rudelson, M.: Random vectors in the isotropic position. J. Funct. Anal. **64**, 60–72 (1999)
24. Steinwart, I., Scovel, C.: Mercers theorem on general domains: On the interaction between measures, kernels, and RKHSs. Constr. Approx. **35**(3):363–417 (2012)
25. Temlyakov, V.: Sampling discretization error of integral norms for function classes. J. Complex. **54**, (2019)
26. Temlyakov, V.N.: The Marcinkiewicz-type discretization theorems. Constr. Approx. **48**(2), 337–369 (2018)
27. Temlyakov, V. N.: On optimal recovery in L_2. J. Complex. **65**, (2021)
28. Tropp, J.: User-friendly tail bounds for sums of random matrices. Found. Comp. Math. **12**(4), 389–434 (2011)
29. Ullrich, M.: On the worst-case error of least squares algorithms for L2-approximation with high probability. Journal of Complexity **60**,(2020)
30. Wasilkowski, G.-W.: Some nonlinear problems are as easy as the approximation problem. Comput. Math. Appl. **10**(4-5), 351–363 (1985)
31. Wasilkowski, G.W., Woźniakowski, H.: On the power of standard information for weighted approximation. Found. Comput. Math. **1**(4), 417–434 (2001)