

Worst-case recovery guarantees for least squares approximation using random samples

Lutz Kämmerer*, Tino Ullrich, Toni Volkmer

TU Chemnitz, Faculty of Mathematics, 09107 Chemnitz, Germany

April 5, 2021

Abstract

We construct a least squares approximation method for the recovery of complex-valued functions from a reproducing kernel Hilbert space on $D \subset \mathbb{R}^d$. The nodes are drawn at random for the whole class of functions and the error is measured in $L_2(D, \varrho_D)$. We prove worst-case recovery guarantees by explicitly controlling all the involved constants. This leads to new preasymptotic recovery bounds with high probability for the error of *Hyperbolic Fourier Regression* on multivariate data. In addition, we further investigate its counterpart *Hyperbolic Wavelet Regression* also based on least-squares to recover non-periodic functions from random samples. Finally, we reconsider the analysis of a cubature method based on plain random points with optimal weights and reveal near-optimal worst-case error bounds with high probability. It turns out that this simple method can compete with the quasi-Monte Carlo methods in the literature which are based on lattices and digital nets.

Keywords and phrases : least squares approximation, random sampling, quadrature, sampling recovery, hyperbolic wavelet regression

2010 AMS Mathematics Subject Classification : 41A10, 41A25, 41A55, 41A60, 41A63, 42A10, 65D32, 65Txx, 68Q25, 68W40, 94A20

1 Introduction

We consider the problem of learning complex-valued multivariate functions on a domain $D \subset \mathbb{R}^d$ from function samples on the set of nodes $\mathbf{X} := \{\mathbf{x}^1, \dots, \mathbf{x}^n\} \subset D$. The functions are modeled as elements from some reproducing kernel Hilbert space $H(K)$ with kernel $K: D \times D \rightarrow \mathbb{C}$. The nodes \mathbf{X} are drawn independently according to a tailored probability measure depending on the spectral properties of the embedding of $H(K)$ in $L_2(D, \varrho_D)$, where the error is measured. Our main focus in this paper is on worst-case recovery guarantees. In fact, we aim at recovering all $f \in H(K)$ simultaneously from sampled values at the sampling nodes in \mathbf{X} with high probability. To be more precise, we implement algorithms and provide error bounds to control the worst-case error

$$\sup_{\|f\|_{H(K)} \leq 1} \|f - S_{\mathbf{X}}^m f\|_{L_2(D, \varrho_D)},$$

*Corresponding author: kaemmerer@mathematik.tu-chemnitz.de

where $S_{\mathbf{X}}^m$ is the fixed recovery operator. In contrast to that, the problem of reconstructing an individual function from random samples has been considered by several authors in the literature, e.g. Smale, Zhou [59], Bohn [5, 6], Bohn, Griebel [7], Cohen, Davenport, Leviatan [13], Chkifa, Migliorati, Nobile, Tempone [10], Cohen, Migliorati [15], and many others to mention just a few.

Let us emphasize, that we do not develop a Monte Carlo method here. It is rather the use of “random information” which gained substantial interest in the Information Based Complexity (IBC) community and in the field of Compressed Sensing, see the recent survey [27] and [23]. We construct a recovery operator $S_{\mathbf{X}}^m$ which computes a best least squares fit $S_{\mathbf{X}}^m f$ to the given data $(f(\mathbf{x}^1), \dots, f(\mathbf{x}^n))^{\top}$ from the finite-dimensional space spanned by the first $m - 1$ singular vectors of the embedding

$$\text{Id} : H(K) \rightarrow L_2(D, \varrho_D). \quad (1.1)$$

The right singular vectors $e_1(\cdot), e_2(\cdot), \dots$ of this embedding are arranged according to their importance, i.e., with respect to the non-increasing rearrangement of the singular values $\sigma_1 \geq \sigma_2 \geq \dots > 0$.

The investigations in this paper are inspired by the recent results by Krieg and M. Ullrich [33], which triggered substantial progress in the field. See also the discussion below in Remark 5.9 and [44, Sect. 7]. In this paper we extend and improve the results from [33] in several directions. In particular, we investigate and implement a least squares regression algorithm under weaker conditions and give practically useful parameter choices which lead to a controlled failure probability and explicit error bounds.

A typical error bound relates the worst-case recovery error to the sequence of singular numbers $(\sigma_k)_{k \in \mathbb{N}}$ of the embedding (1.1) which represent the approximation numbers or linear widths. One main contribution of this paper is the following general bound, where all constants are determined precisely under mild conditions. Recall that $(e_k(\cdot))_{k \in \mathbb{N}}$ denotes the sequence of right singular vectors of the embedding (1.1), i.e., the eigenfunctions of $\text{Id}^* \circ \text{Id} : H(K) \rightarrow H(K)$, and $\sigma_1 \geq \sigma_2 \geq \dots > 0$ the corresponding singular numbers.

Theorem 1.1 (cf. Cor. 5.6). *Let $H(K)$ be a separable reproducing kernel Hilbert space of complex-valued functions on a subset $D \subset \mathbb{R}^d$ such that the positive semidefinite kernel $K : D \times D \rightarrow \mathbb{C}$ satisfies $\sup_{\mathbf{x} \in D} K(\mathbf{x}, \mathbf{x}) < \infty$. Let further ϱ_D denote a probability measure on D . Furthermore, for $n \in \mathbb{N}$ and $\delta \in (0, 1/3)$, we define $m \in \mathbb{N}$ such that*

$$N(m) := \sup_{\mathbf{x} \in D} \sum_{k=1}^{m-1} \sigma_k^{-2} |e_k(\mathbf{x})|^2 \leq \frac{n}{48(\sqrt{2} \log(2n) - \log \delta)}$$

holds. Then the random reconstruction operator $S_{\mathbf{X}}^m$ (see Algorithm 1 below), which uses samples on the n i.i.d. (according to ϱ_D) drawn nodes in \mathbf{X} , satisfies

$$\mathbb{P} \left(\sup_{\|f\|_{H(K)} \leq 1} \|f - S_{\mathbf{X}}^m f\|_{L_2(D, \varrho_D)}^2 \leq \frac{29}{\delta} \max \left\{ \sigma_m^2, \frac{\log(8n)}{n} T(m) \right\} \right) \geq 1 - 3\delta,$$

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

The occurrence of the fundamental quantity $N(m)$ is certainly not a surprise. It is also known as spectral function (see [25] and the references therein) and in case of orthogonal polynomials it is the inverse of the infimum of the Christoffel function (cf. e.g. [22]). It

represents a well-known ingredient for inequalities related to sampling and discretization, see for instance Gröchenig and Bass [1], Gröchenig [24, 25], Temlyakov [61, 62], and Temlyakov et al. [19]. If, for instance, $N(m) \in \mathcal{O}(m)$ holds, we achieve near optimal error bounds with respect to the number of used sampling values n . Note that by a straight-forward computation we also have $T(m) \leq 2 \sum_{k=m/2}^{\infty} \sigma_k^2 N(4k)/k$. Hence, if $N(m) \in \mathcal{O}(m)$, the bound in Theorem 1.1 is upper bounded by

$$\sup_{\|f\|_{H(K)} \leq 1} \|f - S_{\mathbf{X}}^m f\|_{L_2(D, \varrho_D)}^2 \leq \frac{C_{K, \varrho_D}}{\delta \cdot m} \sum_{k \geq m/2}^{\infty} \sigma_k^2 \quad (1.2)$$

with $m := n/(c_1 \log(n) + c_2 \log(\delta^{-1}))$ and a constant $C_{K, \varrho_D} > 0$ depending on the measure ϱ_D and the kernel K .

In the general case we do not necessarily have $N(m) \in \mathcal{O}(m)$. Here, a technique called “importance sampling”, see e.g. [26, 55, 56, 45], turns out to be useful, see Algorithm 2 below. As proposed in [15] and specified in full detail in [33], one may sample from a reweighted distribution using a specific density ϱ_m , defined in (5.12) below, which is different for any m and depends on the spectral properties of the embedding (1.1). It determines the important “area” to sample. In other words, we incorporate additional knowledge about the spectral properties of our embedding. The underlying recovery operator is still constructive and the determined error bounds hold with high probability. Computing this envelope density (and sample from it) has been studied in [15, Sect. 5]. A refinement of this technique together with Theorem 1.1 above leads to the following precise bounds under even weaker conditions.

Theorem 1.2 (cf. Thm. 5.8). *Let $H(K)$ be a separable reproducing kernel Hilbert space of complex-valued functions on a subset $D \subset \mathbb{R}^d$. Let further ϱ_D denote a non-trivial σ -finite measure on D and we assume that the positive semidefinite kernel $K : D \times D \rightarrow \mathbb{C}$ satisfies*

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

Furthermore, for $n \in \mathbb{N}$ and $\delta \in (0, 1/3)$ we fix

$$m := \left\lfloor \frac{n}{96(\sqrt{2} \log(2n) - \log \delta)} \right\rfloor.$$

Then the random reconstruction operator $\tilde{S}_{\mathbf{X}}^m$ (see Algorithm 2 below), which uses n samples drawn according to a probability measure depending on ϱ_D , m and K satisfies

$$\mathbb{P} \left(\sup_{\|f\|_{H(K)} \leq 1} \|f - \tilde{S}_{\mathbf{X}}^m f\|_{L_2(D, \varrho_D)}^2 \leq \frac{50}{\delta} \max \left\{ \sigma_m^2, \frac{\log(8n)}{n} \sum_{j=m}^{\infty} \sigma_j^2 \right\} \right) \geq 1 - 3\delta. \quad (1.3)$$

By the same reasoning as above we may replace the error bound in (1.3) by (1.2) but this time with a universal (and precisely determined) constant $C > 0$. The result refines the bound in [33] as we give precise constants here in the general situation.

A further application of our least squares method is in the field of numerical integration. Oettershagen [50], Belhadji, Bardenet, Chainais [2], Groechenig [25], Migliorati, Nobile [42], and many others used least squares optimization to design quadrature rules. This results in (complex-valued) weights $\mathbf{q} := (q_1, \dots, q_n)^{\top}$ in a cubature formula, i.e.,

$$\tilde{Q}_{\mathbf{X}}^m f = \mathbf{q}^{\top} \cdot \mathbf{f} = \sum_{j=1}^n q_j f(\mathbf{x}^j) := \int_D \tilde{S}_{\mathbf{X}}^m f d\mu_D,$$

where $\mathbf{f} := (f(\mathbf{x}^j))_{j=1}^n$ and μ_D is the measure for which we want to compute the integral. In our setting, the integration nodes $\mathbf{X} = \{\mathbf{x}^1, \dots, \mathbf{x}^n\}$ are determined once in advance for the whole class. Clearly, the bounds from Theorem 1.1, 1.2 can be literally transferred to control the worst-case integration error

$$\sup_{\|f\|_{H(K)} \leq 1} \left| \int_D f(\mathbf{x}) d\mathbf{x} - \tilde{Q}_{\mathbf{X}}^m f \right|,$$

see Theorems 7.1, 7.2.

As the main example we consider the recovery of functions from Sobolev spaces with mixed smoothness (also known as tensor product Sobolev spaces or hyperbolic cross spaces). This problem has been investigated by many authors in the last 30 years, see [17] and the references therein. The above general bound on the worst case errors can for instance be used for any non-periodic embedding

$$\text{Id} : H_{\text{mix}}^s([0, 1]^d) \rightarrow L_2([0, 1]^d), \quad s > 1/2.$$

The spaces $H_{\text{mix}}^s([0, 1]^d)$ can be represented in various ways as a reproducing kernel Hilbert space satisfying the requirements of the above theorems, see the concrete collection of examples in [3, Section 7.4]. Applying Theorem 1.2 and plugging in well-known upper bounds on the singular numbers we improve on the asymptotic sampling bounds in [17, Sect. 5], see also Dinh Dũng [16] and Byrenheid [8] for the non-periodic situation. In addition, using refined preasymptotic estimates for the $(\sigma_j)_{j \in \mathbb{N}}$ in the non-periodic situation (see [32, Section 4.3]) yields reasonable bounds for sampling numbers in case of small n .

Let us emphasize that the result by Krieg, Ullrich [33] can be considered as a major progress for the research on the complexity of this problem. They disproved Conjecture 5.6.2. in [17] for $p = 2$ and $1/2 < s < (d - 1)/2$. Indeed, the celebrated sparse grid points are outperformed in a certain range for s . This represents a first step towards the solution of [17, Outstanding Open Problem 1.4]. As a consequence of the recent contributions by Nagel, Schäfer, T. Ullrich [44] and Temlyakov [63] based on the groundbreaking solution of the Kadison Singer problem by Marcus, Spielman and Srivastava [41] it is now evident that sparse grid methods are not optimal in the full range of parameters (except maybe in $d = 2$). Still it is worth mentioning that the sparse grids represent the best known deterministic construction what concerns the asymptotic order. Indeed, the guarantees are deterministic and only slightly worse compared to random nodes in the asymptotic regime. However, regarding preasymptotics the random constructions provide substantial advantages.

In this paper, we use the simple least squares algorithm from [14, 15, 5, 33], and we show that using random points makes it also possible to obtain explicit worst-case recovery bounds also for small n . In the periodic setting the analysis benefits from the fact that the underlying eigenvector system is a bounded orthonormal system (BOS), see (2.8), which implies in particular $N(m) \in \mathcal{O}(m)$. In case of the complex Fourier system we have a BOS constant $B = 1$ and obtain dimension-free constants. This allows for a priori estimates on the number of required samples and arithmetic operations in order to ensure accuracy $\varepsilon > 0$ with our concrete algorithm. In particular, we incorporate recent preasymptotic bounds for the singular values $(\sigma_k)_{k \in \mathbb{N}}$, see [35, 36] and [32]. We obtain with probability larger than $1 - 3\delta$ for the periodic mixed smoothness space $H_{\text{mix}}^s(\mathbb{T}^d)$ with $2s > 1 + \log_2 d$ equipped with the $\|\cdot\|_{\#}$ -norm (see (8.4) below) the worst-case bound

$$\sup_{\|f\|_{\#} \leq 1} \|f - S_{\mathbf{X}}^{m,\#} f\|_{L_2(\mathbb{T}^d)}^2 \leq \frac{10s}{\delta(2s - 1 - \log_2 d)} \left(\frac{16}{3m} \right)^{\frac{2s}{1 + \log_2 d}}.$$

The number of samples n scales similarly as after (1.2) with precisely determined absolute constants $c_1, c_2 < 70$, see Corollary 8.4 below.

We also demonstrate in Section 9 that the BOS assumption is not necessary for getting a practical algorithm. Similar as in [5] we use an algorithm called “Hyperbolic Wavelet Regression” and show that it recovers non-periodic functions belonging to a Sobolev space with mixed smoothness H_{mix}^s from the n nodes \mathbf{X} drawn according to the uniform measure with a rate similar as for the periodic case. The proposed approach achieves rates which improve on the bounds in [5] and are only worse by $\log^s n$ in comparison to the optimal rates achieved by hyperbolic wavelet best approximation studied in [20] and [58].

Finally, by our refined analysis we were able to settle a conjecture in [50] that the worst-case integration error for $H_{\text{mix}}^s(\mathbb{T}^d)$ is bounded in order by $n^{-s} \log(n)^{ds}$ with high probability. This conjecture was based on the outcome of several numerical experiments (described in [50]) where the worst-case error has been simulated using the RKHS Riesz representer. It is remarkable in two respects. First, it is possible to benefit from higher order smoothness although we use plain random points. And second, this simple method can compete with most of the quasi-Monte Carlo methods based on lattices and digital nets studied in the literature, see [21, pp. 195, 247]. Moreover, if $s < (d - 1)/2$ we get a better asymptotic rate than sparse grid integration which is shown to be of order $n^{-s} \log(n)^{(d-1)(s+1/2)}$, see [18].

We practically verify our theoretical findings with several numerical experiments. There, we compare the recovery error for the least squares regression method $S_{\mathbf{X}}^m$ to the optimal error given by the projection on the eigenvector space. We also study a non-periodic regime, where we randomly sample points according to the Chebyshev measure. Algorithmically, the coefficients $\mathbf{c} := (c_k)_{k=1}^{m-1} \in \mathbb{C}^{m-1}$ of the approximation $S_{\mathbf{X}}^m := \sum_{k=1}^{m-1} c_k \sigma_k^{-1} e_k$ can be obtained by computing the least squares solution of the (over-determined) linear system of equations $\mathbf{L}_m \mathbf{c} = (f(\mathbf{x}^j))_{j=1}^n$, where $\mathbf{L}_m := (\sigma_k^{-1} e_k(\mathbf{x}^j))_{j=1; k=1}^{n; m-1} \in \mathbb{C}^{n \times (m-1)}$. In order to solve this linear system of equations, one can apply a standard conjugate gradient type iterative algorithm, e.g., LSQR [52]. The corresponding arithmetic costs are bounded from above by $C R m n < C R n^2$, where $C > 0$ is an absolute constant and R the number of iterations which is rather small due to the well-conditioned least squares matrices.

Outline. In Section 2 we describe the setting in which we want to perform the worst-case analysis. There we use the framework of reproducing kernel Hilbert spaces of complex-valued functions. Section 3 is devoted to the least squares algorithm, where the worst-case analysis is given in Sections 5, 8, and 9. In the first place, we present the general results in Section 5. Section 8 considers the particular case of hyperbolic Fourier regression. In Section 9 we investigate the particular case of non-periodic functions with a bounded mixed derivative and their recovery via hyperbolic wavelet regression. The main tools from probability theory, like concentration inequalities for spectral norms and Rudelson’s lemma, are provided in Section 4. The analysis of the recovery of individual functions (Monte Carlo) is given in Section 6. Consequences for optimally weighted numerical integration based on plain random points are given in Section 7 and Section 8.2. Finally, the numerical experiments are shown and discussed in Section 10.

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{C} the complex numbers. If not indicated otherwise the symbol \log denotes the natural logarithm. \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. The spectral norm (i.e. the largest singular value) of matrices \mathbf{L} is denoted by $\|\mathbf{L}\|$ or $\|\mathbf{L}\|_{2 \rightarrow 2}$. Vectors and matrices

are usually typesetted bold with $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ or \mathbb{C}^n . For $0 < p \leq \infty$ 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 $p = \infty$. If $T : X \rightarrow Y$ is a continuous operator we write $\|T : X \rightarrow Y\|$ for its operator (quasi-)norm. For two sequences $(a_n)_{n=1}^\infty, (b_n)_{n=1}^\infty \subset \mathbb{R}$ we write $a_n \lesssim b_n$ if there exists a constant $c > 0$ such that $a_n \leq c b_n$ for all n . We will write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. We write $f \in \mathcal{O}(g)$ for non-negative functions f, g if there is a constant $c > 0$ such that $f \leq cg$. D denotes a subset of \mathbb{R}^d and $\ell_\infty(D)$ the set of bounded functions on D with $\|\cdot\|_{\ell_\infty(D)}$ the supremum norm.

2 Reproducing kernel Hilbert spaces

We will work in the framework of reproducing kernel Hilbert spaces. The relevant theoretical background can be found in [3, Chapt. 1] and [11, Chapt. 4]. 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 $K(\mathbf{x}, \mathbf{y})$ on $D \times D$. The crucial property is the identity

$$f(\mathbf{x}) = \langle f, K(\cdot, \mathbf{x}) \rangle_{H(K)} \quad (2.1)$$

for all $\mathbf{x} \in D$. It ensures that point evaluations are continuous functionals on $H(K)$. We will use the notation from [11, Chapt. 4]. In the framework of this paper, the finite trace of the kernel

$$\int_D K(\mathbf{x}, \mathbf{x}) \varrho_D(d\mathbf{x}) < \infty \quad (2.2)$$

or its boundedness

$$\|K\|_\infty := \sup_{\mathbf{x} \in D} \sqrt{K(\mathbf{x}, \mathbf{x})} < \infty \quad (2.3)$$

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

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

Note that we do not need the measure ϱ_D for this embedding.

The embedding operator

$$\text{Id} : H(K) \rightarrow L_2(D, \varrho_D)$$

is compact under the integrability condition (2.2), which we always assume from now on. We additionally assume that $H(K)$ is at least infinite dimensional. However, we do not assume the separability of $H(K)$ here. Due to the compactness of Id the operator $\text{Id}^* \circ \text{Id}$ provides an at most countable system of strictly positive eigenvalues $(\lambda_j)_{j \in \mathbb{N}}$. These eigenvalues are summable as a consequence of (2.2) and (2.5), (2.6) below, such that the singular numbers $(\sigma_j)_{j \in \mathbb{N}}$ belong to ℓ_2 . Indeed, let Id^* be defined in the usual way as

$$\langle \text{Id}(f), g \rangle_{L_2} = \langle f, \text{Id}^*(g) \rangle_{H(K)}.$$

Then $W_{\varrho_D} := \text{Id}^* \circ \text{Id} : H(K) \rightarrow H(K)$ is non-negative definite, self-adjoint and compact. Let $(\lambda_j, e_j)_{j \in \mathbb{N}}$ denote the eigenpairs of W_{ϱ_D} , where $(e_j)_{j \in \mathbb{N}} \subset H(K)$ is an orthonormal system of eigenvectors, and $(\lambda_j)_{j \in \mathbb{N}}$ the corresponding positive eigenvalues. In fact, $W_{\varrho_D} e_j = \lambda_j e_j$ and $\langle e_j, e_k \rangle_{H(K)} = \delta_{j,k}$. The sequence of positive eigenvalues are arranged in non-increasing order, i.e.,

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

Note that we have by Bessel's inequality

$$\|f\|_{H(K)}^2 \geq \sum_{k=1}^{\infty} |\langle f, e_k \rangle_{H(K)}|^2. \quad (2.4)$$

Let us point out, that by (2.4), the function

$$\mathbf{x} \mapsto \sum_{k=1}^{\infty} |e_k(\mathbf{x})|^2 \leq K(\mathbf{x}, \mathbf{x}) \quad (2.5)$$

exists pointwise in \mathbb{C} . This implies

$$\sum_{k=1}^{\infty} \lambda_k \leq \int_D K(\mathbf{x}, \mathbf{x}) \varrho_D(d\mathbf{x}) \quad (2.6)$$

and we get by $\lambda_j = \sigma_j^2$ that $\text{Id} : H(K) \rightarrow L_2(D, \varrho_D)$ is a Hilbert-Schmidt operator if (2.2) holds. We will restrict to the situation where we have equality in (2.6). This can be achieved by posing additional assumptions, namely that $H(K)$ is separable and ϱ_D is σ -finite, see [11, Thm. 4.27]. It further holds that

$$\langle e_j, e_k \rangle_{L_2} = \langle \text{Id}(e_j), \text{Id}(e_k) \rangle_{L_2} = \langle We_j, e_k \rangle_{H(K)} = \lambda_j \langle e_j, e_k \rangle_{H(K)} = \lambda_j \delta_{j,k}.$$

Hence, $(e_j)_{j \in \mathbb{N}}$ is also orthogonal in $L_2(D, \varrho_D)$ and $\|e_j\|_2 = \sqrt{\lambda_j} =: \sigma_j$. We define the orthonormal system $(\eta_j)_{j \in \mathbb{N}} := (\lambda_j^{-1/2} e_j)_{j \in \mathbb{N}}$ in $L_2(D, \varrho_D)$.

For our subsequent analysis the quantity

$$N(m) := \sup_{\mathbf{x} \in D} \sum_{k=1}^{m-1} |\eta_k(\mathbf{x})|^2 \quad (2.7)$$

plays a fundamental role. We often need the related quantity $T(m) := \sup_{\mathbf{x} \in D} \sum_{k=m}^{\infty} |\eta_k(\mathbf{x})|^2$ which can be estimated by $T(m) \leq 2 \sum_{k=m/2}^{\infty} \sigma_k^2 N(4k)/k$. The first one is sometimes called “spectral function”, see [25] and the references therein. Clearly, by (2.6) $N(m)$ and $T(m)$ are well-defined if the kernel is bounded, i.e., if (2.3) is assumed. In fact, $T(m)$ is bounded by $\|K\|_{\infty}$. It may happen that the system $(\eta_k)_{k \in \mathbb{N}}$ is a uniformly $\ell_{\infty}(D)$ bounded orthonormal system (BOS), i.e., we have for all $k \in \mathbb{N}$

$$\|\eta_k\|_{\ell_{\infty}(D)} \leq B. \quad (2.8)$$

Let us call B the BOS constant of the system. In this case we have $N(m) \leq (m-1)B^2 \in \mathcal{O}(m)$ and $T(m) \leq B^2 \sum_{k=m}^{\infty} \lambda_k$.

Remark 2.1. We would like to point out an issue concerning the embedding operator $\text{Id} : H(K) \rightarrow L_2(D, \varrho_D)$ defined above. As discussed in [11, p. 127] this embedding operator is in general *not* injective as it maps a function to an equivalence class. As a consequence, the system of eigenvectors $(e_j)_{j \in \mathbb{N}}$ may not be a basis in $H(K)$ (note that $H(K)$ may not even be separable). However, there are conditions which ensure that the orthonormal system $(e_j)_{j \in \mathbb{N}}$ is an orthonormal basis in $H(K)$, see [11, 4.5] and [11, Ex. 4.6, p. 163], which is related to Mercer's theorem [11, Thm. 4.49]. Indeed, if we additionally assume that the kernel $K(\cdot, \cdot)$ is bounded and continuous on $D \times D$ (for a domain $D \subset \mathbb{R}^d$), then $H(K)$ is separable and consists of continuous functions, see [3, Thms. 16, 17]. If we finally assume that the measure

ϱ_D is a Borel measure with full support then $(e_j)_{j \in \mathbb{N}}$ is a complete orthonormal system in $H(K)$. In this case we have the pointwise identity

$$K(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{\infty} \overline{e_j(\mathbf{y})} e_j(\mathbf{x}), \quad \mathbf{x}, \mathbf{y} \in D, \quad (2.9)$$

as well as equality signs in (2.4), (2.5) and (2.6), see for instance [3, Cor. 4]. Let us emphasize, that a Mercer kernel $K(\cdot, \cdot)$, which is a continuous kernel on $D \times D$ with a compact domain $D \subset \mathbb{R}^d$ satisfies all these conditions, see [11, Thm. 4.49]. In this case, we even have (2.9) with absolute and uniform convergence on $D \times D$. Let us point out that, to our surprise, Theorem 1.2 (Theorem 5.8) holds already true under the finite trace condition (2.2) if $H(K)$ is separable and ϱ_D is σ -finite. We do not have to assume continuity of the kernel. Note that the finite trace condition is natural in this context as [28] shows.

3 Least squares regression

Our algorithm essentially boils down to the solution of an over-determined system

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

where $\mathbf{L} \in \mathbb{C}^{n \times m}$ is a matrix with $n > m$. 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 $\|\mathbf{f} - \mathbf{L} \mathbf{c}\|_2$. Multiplying the system with \mathbf{L}^* gives

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

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

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

From the fact that the singular values of \mathbf{L} are bounded away from zero we get the following quantitative bound on the spectral norm of the Moore-Penrose inverse $(\mathbf{L}^* \mathbf{L})^{-1} \mathbf{L}^*$.

Proposition 3.1. *Let $\mathbf{L} \in \mathbb{C}^{n \times m}$ be a matrix with $m < n$ with full rank and singular values $\tau_1, \dots, \tau_m > 0$ arranged in non-increasing order.*

- (i) *Then also the matrix $(\mathbf{L}^* \mathbf{L})^{-1} \mathbf{L}^*$ has full rank and singular values $\tau_m^{-1}, \dots, \tau_1^{-1}$ (arranged in non-increasing order).*
- (ii) *In particular, it holds that*

$$(\mathbf{L}^* \mathbf{L})^{-1} \mathbf{L}^* = \mathbf{V}^* \tilde{\Sigma} \mathbf{U}$$

whenever $\mathbf{L} = \mathbf{U}^ \Sigma \mathbf{V}$, where $\Sigma \in \mathbb{R}^{n \times m}$ is a rectangular matrix only with (τ_1, \dots, τ_m) on the “main diagonal” and orthogonal matrices $\mathbf{U} \in \mathbb{C}^{n \times n}$ and $\mathbf{V} \in \mathbb{C}^{m \times m}$. Here $\tilde{\Sigma} \in \mathbb{R}^{m \times n}$ denotes the matrix with $(\tau_1^{-1}, \dots, \tau_m^{-1})$ on the “main diagonal”.*

- (iii) *The operator norm $\|(\mathbf{L}^* \mathbf{L})^{-1} \mathbf{L}^*\|$ can be controlled as follows*

$$\|(\mathbf{L}^* \mathbf{L})^{-1} \mathbf{L}^*\| \leq \tau_m^{-1}.$$

For function recovery we will use the following matrix

$$\mathbf{L}_m := \mathbf{L}_m(\mathbf{X}) = \begin{pmatrix} \eta_1(\mathbf{x}^1) & \eta_2(\mathbf{x}^1) & \cdots & \eta_{m-1}(\mathbf{x}^1) \\ \vdots & \vdots & & \vdots \\ \eta_1(\mathbf{x}^n) & \eta_2(\mathbf{x}^n) & \cdots & \eta_{m-1}(\mathbf{x}^n) \end{pmatrix}, \quad (3.1)$$

for $\mathbf{X} = \{\mathbf{x}^1, \dots, \mathbf{x}^n\} \subset D$ of distinct sampling nodes and the system $(\eta_k)_{k \in \mathbb{N}} := (\lambda_k^{-1/2} e_k)_{k \in \mathbb{N}}$. Below we will see that this matrix behaves well with high probability if n is large enough and the nodes in \mathbf{X} are chosen independently and identically ϱ_D -distributed from D .

Algorithm 1 Least squares regression.

Input:	$\mathbf{X} = \{\mathbf{x}^1, \dots, \mathbf{x}^n\} \subset D$	set of distinct sampling nodes,
	$\mathbf{f} = (f(\mathbf{x}^1), \dots, f(\mathbf{x}^n))^\top$	samples of f evaluated at the nodes from \mathbf{X} ,
	$m \in \mathbb{N}$	$m < n$ such that the matrix $\mathbf{L}_m := \mathbf{L}_m(\mathbf{X})$ from (3.1) has full (column) rank.

Solve the over-determined linear system

$$\mathbf{L}_m (c_1, \dots, c_{m-1})^\top = \mathbf{f}$$

via least squares (e.g. directly or via the LSQR algorithm [52]), i.e., compute

$$(c_1, \dots, c_{m-1})^\top := (\mathbf{L}_m^* \mathbf{L}_m)^{-1} \mathbf{L}_m^* \mathbf{f}.$$

Output: $\mathbf{c} = (c_1, \dots, c_{m-1})^\top \in \mathbb{C}^{m-1}$ coefficients of the approximant $S_{\mathbf{X}}^m f := \sum_{k=1}^{m-1} c_k \eta_k$.

Using Algorithm 1, we compute the coefficients c_k , $k = 1, \dots, m-1$, of the approximant

$$S_{\mathbf{X}}^m f := \sum_{k=1}^{m-1} c_k \eta_k. \quad (3.2)$$

Note that the mapping $f \mapsto S_{\mathbf{X}}^m f$ is linear for a fixed set of sampling nodes $\mathbf{X} \subset D$.

4 Concentration inequalities

We will consider complex-valued random variables X and random vectors (X_1, \dots, X_N) on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. As usual we will denote with $\mathbb{E}X$ the expectation of X . With $\mathbb{P}(A|B)$ and $\mathbb{E}(X|B)$ we denote the conditional probability

$$\mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

and the conditional expectation

$$\mathbb{E}(X|B) = \frac{\mathbb{E}(\chi_B \cdot X)}{\mathbb{P}(B)}, \quad (4.1)$$

where $\chi_B : \Omega \rightarrow \{0, 1\}$ is the indicator function on B .

Let us start with the classical Markov inequality. If Z is a random variable then

$$\mathbb{P}(|Z| > t) \leq \frac{\mathbb{E}|Z|}{t}, \quad t > 0.$$

Of course, there is also a version involving conditional probability and expectation. In fact,

$$\mathbb{P}(|Z| > t \mid B) \leq \frac{\mathbb{E}(|Z| \mid B)}{t}, \quad t > 0. \quad (4.2)$$

Let us state concentration inequalities for the norm of sums of complex rank-1 matrices. For the first result we refer to Oliveira [51]. We will need the following notational convention: For a complex (column) vector $\mathbf{y} \in \mathbb{C}^N$ (or ℓ_2) we will often use the tensor notation for the matrix

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

Proposition 4.1. *Let $\mathbf{y}^i, i = 1, \dots, n$, be i.i.d. copies of a random vector $\mathbf{y} \in \mathbb{C}^N$ such that $\|\mathbf{y}^i\|_2 \leq M$ almost surely. Let further $\mathbb{E}(\mathbf{y}^i \otimes \mathbf{y}^i) = \Lambda \in \mathbb{C}^{N \times N}$ and $0 < t < 1$. Then it holds*

$$\mathbb{P}\left(\left\|\frac{1}{n} \sum_{i=1}^n \mathbf{y}^i \otimes \mathbf{y}^i - \Lambda\right\| > t\right) \leq (2 \min(n, N))^{\sqrt{2}} \exp\left(-\frac{nt^2}{12M^2}\right).$$

Proof. In order to show the probability estimate, we refer to the proof of [51, Lem. 1] and observe

$$\mathbb{P}\left(\left\|\frac{1}{n} \sum_{i=1}^n \mathbf{y}^i \otimes \mathbf{y}^i - \Lambda\right\| > t\right) \leq (2 \min(n, N))^{\frac{1}{1-2M^2s/n}} \exp\left(-st + \frac{2M^2s^2}{n-2M^2s}\right)$$

for $0 \leq s \leq n/(2M^2)$. Since we restrict $0 < t < 1$, the choice $s = (4 + 2\sqrt{2})^{-1}nt/M^2$ yields

$$(2 \min(n, N))^{\frac{1}{1-2M^2s/n}} \exp\left(-st + \frac{2M^2s^2}{n-2M^2s}\right) = (2 \min(n, N))^{\sqrt{2}} \exp\left(-\frac{nt^2}{(6 + 4\sqrt{2})M^2}\right)$$

and, finally, the assertion holds. \square

Remark 4.2. A slightly stronger version for the case of real matrices can be found in Cohen, Davenport, Leviatan [13] (see also the correction in [14]). For $\mathbf{y}^i, i = 1, \dots, n$, i.i.d. copies of a random vector $\mathbf{y} \in \mathbb{R}^N$ sampled from a bounded orthonormal system, one obtains the concentration inequality

$$\mathbb{P}\left(\left\|\frac{1}{n} \sum_{i=1}^n \mathbf{y}^i \otimes \mathbf{y}^i - \mathbf{I}\right\| > t\right) \leq 2N \exp(-c_t n/M^2),$$

where $c_t = (1+t)(\ln(1+t)) - t$. This leads to improved constants for the case of real matrices.

The following result goes back to Lust-Piquard, Pisier [40], and Rudelson [57]. The complex version with precise constants can be found in Rauhut [54, Cor. 6.20].

Proposition 4.3. *Let $\mathbf{y}^i \in \mathbb{C}^N$ (or ℓ_2), $i = 1, \dots, n$, and ε_i independent Rademacher variables taking values ± 1 with equal probability. Then*

$$\mathbb{E}_\varepsilon \left\| \sum_{i=1}^n \varepsilon_i \mathbf{y}^i \otimes \mathbf{y}^i \right\| \leq C_R \sqrt{\log(8 \min\{n, N\})} \cdot \sqrt{\left\| \sum_{i=1}^n \mathbf{y}^i \otimes \mathbf{y}^i \right\|} \cdot \max_{i=1, \dots, n} \|\mathbf{y}^i\|_2, \quad (4.3)$$

with

$$C_R = \sqrt{2} + \frac{1}{4\sqrt{2 \log(8)}} \in [1.53, 1.54]. \quad (4.4)$$

Remark 4.4. The result is proved for complex (finite) matrices. Note that the factor $\sqrt{\log(8\min\{n, N\})}$ is already an upper bound for $\sqrt{\log(8r)}$, where r is the rank of the matrix $\sum_i \mathbf{y}^i \otimes \mathbf{y}^i$. The proof of Proposition 4.3 with the precise constant is based on [54, Lem. 6.18] which itself is based on a non-commutative Khintchine inequality, see [54, 6.5]. This technique allows for controlling all the involved constants. Let us comment on the situation $N = \infty$, i.e., $\mathbf{y}^j \in \ell_2$, where this inequality keeps valid with the factor $\sqrt{\log(8n)}$. In fact, if the matrices \mathbf{B}_j in [54, Thm. 6.14] are replaced by rank-1-operators $\mathbf{B}_j: \ell_2 \rightarrow \ell_2$ of type $\mathbf{B}_j = \mathbf{y}^j \otimes \mathbf{y}^j$ with $\|\mathbf{y}^j\|_2 < \infty$ then all the arguments keep valid and an ℓ_2 -version of this non-commutative Khintchine inequality is available. This implies an ℓ_2 -version of [54, Lem. 6.18] which reads as follows: Let $\mathbf{y}^j \in \ell_2$, $j = 1, \dots, n$, and $p \geq 2$. Then

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

Since we control the moments of the random variable representing the norm on the left-hand side we are now able to derive a concentration inequality by standard arguments ([54, Prop. 6.5]). This concentration inequality then easily implies the ℓ_2 -version of (4.3).

As a consequence of this result we obtain the following deviation inequality in the mean which will be sufficient for our purpose.

Corollary 4.5. *Let \mathbf{y}^i , $i = 1, \dots, n$, be i.i.d. random vectors from \mathbb{C}^N or ℓ_2 with $\|\mathbf{y}^i\|_2 \leq M$ almost surely. Let further $\mathbf{\Lambda} = \mathbb{E}(\mathbf{y}^i \otimes \mathbf{y}^i)$. Then with $N > n$ we obtain*

$$\mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{y}^i \otimes \mathbf{y}^i - \mathbf{\Lambda} \right\| \leq 4C_R^2 \frac{\log(8n)}{n} M^2 + 2C_R \sqrt{\frac{\log(8n)}{n}} M \sqrt{\|\mathbf{\Lambda}\|}.$$

Proof. By well-known symmetrization technique (see [23, Lem. 8.4]), Proposition 4.3, and the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} F := \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{y}^i \otimes \mathbf{y}^i - \mathbf{\Lambda} \right\| &\leq 2\mathbb{E}_{\mathbf{y}} \mathbb{E}_\varepsilon \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \mathbf{y}^i \otimes \mathbf{y}^i \right\| \\ &\leq 2C_R \frac{\sqrt{\log(8n)}}{n} \left(\mathbb{E} \max_{i=1, \dots, n} \|\mathbf{y}^i\|_2^2 \right)^{1/2} \left(\mathbb{E} \left\| \sum_{i=1}^n \mathbf{y}^i \otimes \mathbf{y}^i \right\| \right)^{1/2} \\ &\leq \underbrace{2C_R \frac{\sqrt{\log(8n)}}{\sqrt{n}} M}_{=:a} \underbrace{\left(\mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{y}^i \otimes \mathbf{y}^i - \mathbf{\Lambda} \right\| + \underbrace{\|\mathbf{\Lambda}\|}_{=:b} \right)^{1/2}}_{=:F}. \end{aligned}$$

Hence, we get $F^2 \leq a^2(F + b)$ and we solve this inequality with respect to F , which yields

$$0 \leq F \leq \frac{a^2}{2} + \sqrt{\frac{a^4}{4} + a^2b} \leq a^2 + a\sqrt{b}$$

and this corresponds to the assertion. \square

5 Worst-case errors for least-squares regression

5.1 Random matrices from sampled orthonormal systems

Let us start with a concentration inequality for the spectral norm of a matrix of type (3.1). It turns out that 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 the elements of the set $\mathbf{X} \subset D$ of sampling nodes are drawn i.i.d. at random according to ϱ_D . We will find below that the eigenvalues of

$$\mathbf{H}_m := \mathbf{H}_m(\mathbf{X}) = \frac{1}{n} \mathbf{L}_m^* \mathbf{L}_m \in \mathbb{C}^{(m-1) \times (m-1)}$$

are bounded away from zero with high probability if m is small enough compared to n . We speak of an ‘‘oversampling factor’’ n/m . In case of a bounded orthonormal system with BOS constant B , see (2.8), it will turn out that a logarithmic oversampling is sufficient, see (5.2) below. Note that the boundedness constant B may also depend on the underlying spatial dimension d . However, if for instance the complex Fourier system $\{\exp(2\pi i \mathbf{k} \cdot \mathbf{x}) : \mathbf{k} \in \mathbb{Z}^d\}$ is considered, we are in the comfortable situation that $B = 1$.

Proposition 5.1. *Let $n, m \in \mathbb{N}$, $m \geq 2$. Let further $\{\eta_1(\cdot), \eta_2(\cdot), \eta_3(\cdot), \dots\}$ be the orthonormal system in $L_2(D, \varrho_D)$ induced by the kernel K and the n sampling nodes in \mathbf{X} be drawn i.i.d. at random according to ϱ_D . Then it holds for $0 < t < 1$ that*

$$\mathbb{P}(\|\mathbf{H}_m - \mathbf{I}_m\| > t) \leq (2n)^{\sqrt{2}} \exp\left(-\frac{nt^2}{12 \cdot N(m)}\right),$$

where $N(m)$ is defined in (2.7) and $\mathbf{I}_m = \text{diag}(\mathbf{1}) \in \{0, 1\}^{(m-1) \times (m-1)}$.

Proof. We set $\mathbf{y}^i := (\eta_1(\mathbf{x}^i), \dots, \eta_{m-1}(\mathbf{x}^i))^*$, $i = 1, \dots, n$, and observe

$$\mathbf{H}_m := \mathbf{H}_m(\mathbf{X}) = \frac{1}{n} \mathbf{L}_m^* \mathbf{L}_m = \frac{1}{n} \sum_{i=1}^n \mathbf{y}^i \otimes \mathbf{y}^i.$$

Moreover, due to the fact that we have an orthonormal system $(\eta_k)_{k \in \mathbb{N}}$, we obtain that $\mathbb{E}(\mathbf{H}_m) = \mathbf{I}_m$. The result follows by noting that $M^2 \leq N(m)$ in Proposition 4.1. \square

Remark 5.2. From this proposition we immediately obtain that the matrix $\mathbf{H}_m \in \mathbb{C}^{(m-1) \times (m-1)}$ has only eigenvalues larger than $t := 1/2$ with probability at least $1 - \delta$ if

$$N(m) \leq \frac{n}{48(\sqrt{2} \log(2n) - \log \delta)}. \quad (5.1)$$

Hence, in case of a bounded orthonormal system with BOS constant $B > 0$, see (2.8), we may choose

$$m \leq \kappa_{\delta, B} \frac{n}{\log(2n)} \quad (5.2)$$

with $\kappa_{\delta, B} := (\log(1/\delta) + \sqrt{2})^{-1} B^{-2}/48$.

From Proposition 5.1 we get that all $m-1$ singular values $\tau_1, \dots, \tau_{m-1}$ of \mathbf{L}_m from (3.1) are all not smaller than $\sqrt{n/2}$ and not larger than $\sqrt{3n/2}$ with probability at least $1 - \delta$ if m is chosen such that (5.1) holds. In terms of Proposition 3.1 this means that $\tau_1, \dots, \tau_{m-1} \geq \sqrt{n/2}$. This leads to an upper bound on the norm of the Moore-Penrose inverse that is required for the least squares algorithm.

Proposition 5.3. Let $\{\eta_1(\cdot), \eta_2(\cdot), \eta_3(\cdot), \dots\}$ be the orthonormal system in $L_2(D, \varrho_D)$ induced by the kernel K . Let further $m, n \in \mathbb{N}$, $m \geq 2$, and $0 < \delta < 1$ be chosen such that they satisfy (5.1). Then the random matrix \mathbf{L}_m from (3.1) satisfies

$$\|(\mathbf{L}_m^* \mathbf{L}_m)^{-1} \mathbf{L}_m^*\| \leq \sqrt{\frac{2}{n}}$$

with probability at least $1 - \delta$.

In addition to the matrix \mathbf{L}_m , we need to consider a second linear operator that is defined using sampling values of the eigenfunctions e_j . The importance of this operator has been pointed out in [33], where strong results on the concentration of infinite dimensional random matrices have been used. Since we only need the expectation of the norm, we only use Rudelson's lemma, see Proposition 4.3, and a symmetrization technique. This allows us to control the constants.

Proposition 5.4. Let $\mathbf{X} = \{\mathbf{x}^1, \dots, \mathbf{x}^n\} \subset D$ be a set of n sampling nodes drawn uniformly and i.i.d. at random according to ϱ_D , and consider the n i.i.d. random sequences

$$\mathbf{y}^i = (e_m(\mathbf{x}^i), e_{m+1}(\mathbf{x}^i), \dots)^\top, \quad i = 1, \dots, n,$$

together with $T(m) := \sup_{\mathbf{x} \in D} \sum_{k=m}^{\infty} |e_k(\mathbf{x})|^2 < \infty$, see (2.7). Then the operator

$$\Phi_m: \ell_2 \rightarrow \mathbb{R}^n, \quad \mathbf{z} \mapsto \begin{pmatrix} \langle \mathbf{z}, \mathbf{y}^1 \rangle_{\ell_2} \\ \vdots \\ \langle \mathbf{z}, \mathbf{y}^n \rangle_{\ell_2} \end{pmatrix}$$

has expected norm

$$\mathbb{E}(\|\Phi_m\|^2) \leq n \left(\sigma_m^2 + 4 C_R^2 \frac{\log(8n)}{n} T(m) + 2 C_R \sigma_m \sqrt{\frac{\log(8n)}{n} T(m)} \right). \quad (5.3)$$

Proof. Note that $\Phi_m^* \Phi_m = \sum_{i=1}^n \mathbf{y}^i \otimes \mathbf{y}^i$ and

$$\Lambda_m := \mathbb{E}\left(\frac{1}{n} \Phi_m^* \Phi_m\right) = \mathbb{E}(\mathbf{y}^i \otimes \mathbf{y}^i) = \text{diag}(\sigma_m^2, \sigma_{m+1}^2, \dots).$$

This gives

$$\|\Phi_m\|^2 = \|\Phi_m^* \Phi_m\| \leq \|\Phi_m^* \Phi_m - n \Lambda_m\| + n \|\Lambda_m\|.$$

Finally, the bound in (5.3) follows from Corollary 4.5 (see also Remark 4.4 for $N = \infty$), the fact that $\|\Lambda_m\| = \lambda_m = \sigma_m^2$ and $M^2 = T(m)$. \square

5.2 Worst-case errors with high probability

Theorem 5.5. Let $H(K)$ be a separable reproducing kernel Hilbert space on a domain $D \subset \mathbb{R}^d$ with a positive semidefinite kernel $K(\mathbf{x}, \mathbf{y})$ 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 $\text{Id}: H(K) \rightarrow L_2(D, \varrho_D)$ for a probability measure ϱ_D . Let further $0 < \delta < 1$ and $m, n \in \mathbb{N}$, where $m \geq 2$ is

chosen such that (5.1) holds. Drawing the n sampling nodes in \mathbf{X} i.i.d. at random according to ϱ_D , we have for the conditional expectation of the worst-case error

$$\begin{aligned} & \mathbb{E}\left(\sup_{\|f\|_{H(K)} \leq 1} \|f - S_{\mathbf{X}}^m f\|_{L_2(D, \varrho_D)}^2 \mid \|\mathbf{H}_m - \mathbf{I}_m\| \leq 1/2\right) \\ & \leq \frac{1}{1-\delta} \left(3\sigma_m^2 + 8C_R^2 \frac{\log(8n)}{n} T(m) + 4C_R \sigma_m \sqrt{\frac{\log(8n)}{n} T(m)}\right) \quad (5.4) \\ & \leq \frac{3 + 8C_R^2 + 4C_R}{1-\delta} \max\left\{\sigma_m^2, \frac{\log(8n)}{n} T(m)\right\} \end{aligned}$$

with C_R from (4.4).

Proof. Let $f \in H(K)$ such that $\|f\|_{H(K)} \leq 1$. Let further $\mathbf{X} = \{\mathbf{x}^1, \dots, \mathbf{x}^n\}$ be such that $\|\mathbf{H}_m - \mathbf{I}_m\| \leq 1/2$. Using orthogonality and the reproducing property $S_{\mathbf{X}}^m P_{m-1} f = P_{m-1} f$, we estimate

$$\begin{aligned} \|f - S_{\mathbf{X}}^m f\|_{L_2(D, \varrho_D)}^2 &= \|f - P_{m-1} f\|_{L_2(D, \varrho_D)}^2 + \|P_{m-1} f - S_{\mathbf{X}}^m f\|_{L_2(D, \varrho_D)}^2 \\ &\leq \sigma_m^2 + \|S_{\mathbf{X}}^m(P_{m-1} f - f)\|_{L_2(D, \varrho_D)}^2 \\ &= \sigma_m^2 + \left\| (\mathbf{L}_m^* \mathbf{L}_m)^{-1} \mathbf{L}_m^* \left((P_{m-1} f - f)(\mathbf{x}^k) \right)_{k=1}^n \right\|_2^2 \quad (5.5) \\ &\leq \sigma_m^2 + \frac{2}{n} \sum_{k=1}^n \left| (f - P_{m-1} f)(\mathbf{x}^k) \right|^2, \end{aligned}$$

where $P_{m-1} f$ denotes the projection $\sum_{j=1}^{m-1} \langle f, e_j \rangle e_j$ in $H(K)$ yielding $\|f - P_{m-1} f\|_{L_2(D, \varrho_D)} \leq \sigma_m$. Note further, that for any $\mathbf{x} \in D$

$$\begin{aligned} (f - P_{m-1} f)(\mathbf{x}) &= \langle f, K(\cdot, \mathbf{x}) - \sum_{j=1}^{\infty} e_j(\cdot) \overline{e_j(\mathbf{x})} + \sum_{j=1}^{\infty} e_j(\cdot) \overline{e_j(\mathbf{x})} - \sum_{j=1}^{m-1} e_j(\cdot) \overline{e_j(\mathbf{x})} \rangle_{H(K)} \\ &= \sum_{j=m}^{\infty} \langle f, e_j \rangle_{H(K)} e_j(\mathbf{x}) + \langle f, T(\cdot, \mathbf{x}) \rangle_{H(K)}, \end{aligned}$$

where $T(\cdot, \mathbf{x}) = K(\cdot, \mathbf{x}) - \sum_{j=1}^{\infty} e_j(\cdot) \overline{e_j(\mathbf{x})}$ denotes an element in $H(K)$. Its norm is given by

$$\|T(\cdot, \mathbf{x})\|_{H(K)}^2 := \langle T(\cdot, \mathbf{x}), T(\cdot, \mathbf{x}) \rangle_{H(K)} = K(\mathbf{x}, \mathbf{x}) - \sum_{j=1}^{\infty} |e_j(\mathbf{x})|^2. \quad (5.6)$$

This gives

$$\begin{aligned} \left| (f - P_{m-1} f)(\mathbf{x}) \right|^2 &\leq \left| \sum_{j=m}^{\infty} \langle f, e_j \rangle_{H(K)} e_j(\mathbf{x}) \right|^2 + 2\|f\|_{H(K)}^2 \|T(\cdot, \mathbf{x})\|_{H(K)} \sqrt{\sum_{j=m}^{\infty} |e_j(\mathbf{x})|^2} \\ &\quad + \|f\|_{H(K)}^2 \|T(\cdot, \mathbf{x})\|_{H(K)}^2 \\ &\stackrel{(2.5)}{\leq} \left(\sum_{i=m}^{\infty} \overline{\langle f, e_i \rangle_{H(K)} e_i(\mathbf{x})} \right) \left(\sum_{j=m}^{\infty} \langle f, e_j \rangle_{H(K)} e_j(\mathbf{x}) \right) \\ &\quad + \|f\|_{H(K)}^2 \|T(\cdot, \mathbf{x})\|_{H(K)} \left(\|T(\cdot, \mathbf{x})\|_{H(K)} + 2\|K\|_{\infty} \right) \\ &\leq \sum_{i=m}^{\infty} \sum_{j=m}^{\infty} \overline{\langle f, e_i \rangle_{H(K)}} \langle f, e_j \rangle_{H(K)} \overline{e_i(\mathbf{x})} e_j(\mathbf{x}) \\ &\quad + 3\|T(\cdot, \mathbf{x})\|_{H(K)} \|K\|_{\infty} \|f\|_{H(K)}^2. \end{aligned}$$

Returning to (5.5) we estimate

$$\begin{aligned} \sum_{k=1}^n \left| (f - P_{m-1}f)(\mathbf{x}^k) \right|^2 &\leq \|(\langle f, e_j \rangle_{H(K)})_{j \in \mathbb{N}}\|_2^2 \|\Phi_m\|^2 + 3\|K\|_\infty \|f\|_{H(K)}^2 \sum_{k=1}^n \|T(\cdot, \mathbf{x}^k)\| \\ &\leq \|f\|_{H(K)}^2 \left(\|\Phi_m\|^2 + 3\|K\|_\infty \sum_{k=1}^n \|T(\cdot, \mathbf{x}^k)\| \right), \end{aligned} \quad (5.7)$$

where Φ_m denotes the infinite matrix from Proposition 5.4. Note that we used (2.4) in the last but one step. The relation in (5.7) together with (5.5) and $\|f\|_{H(K)} \leq 1$ implies

$$\begin{aligned} \|f - S_{\mathbf{X}}^m f\|_{L_2(D, \varrho_D)}^2 &\leq \sigma_m^2 + \|(\mathbf{L}_m^* \mathbf{L}_m)^{-1} \mathbf{L}_m^*\|^2 \cdot \sum_{k=1}^n \left| (f - P_{m-1}f)(\mathbf{x}^k) \right|^2 \\ &= \sigma_m^2 + \frac{2}{n} \|\Phi_m\|^2 + \frac{6\|K\|_\infty}{n} \sum_{k=1}^n \|T(\cdot, \mathbf{x}^k)\|. \end{aligned}$$

Integrating on both sides yields

$$\begin{aligned} &\int_{\|\mathbf{H}_m - \mathbf{I}_m\| \leq 1/2} \sup_{\|f\|_{H(K)} \leq 1} \|f - S_{\mathbf{X}}^m f\|_{L_2(D, \varrho_D)}^2 \varrho_D^n(d\mathbf{X}) \\ &\leq \sigma_m^2 + \frac{2}{n} \mathbb{E}(\|\Phi_m\|^2) + \frac{6\|K\|_\infty}{n} \sum_{k=1}^n \int_D \|T(\cdot, \mathbf{x}^k)\| \varrho_D(d\mathbf{x}) \\ &= \sigma_m^2 + \frac{2}{n} \mathbb{E}(\|\Phi_m\|^2). \end{aligned} \quad (5.8)$$

Note that the integral on the right-hand side of (5.8) vanishes because of (5.6) and the fact that we have an equality sign in (2.6) due to our assumptions (separability of $H(K)$). This gives

$$0 = \int_D K(\mathbf{x}, \mathbf{x}) \varrho_D(d\mathbf{x}) - \int_D \sum_{j=1}^{\infty} \lambda_j \varrho_D(d\mathbf{x}) = \int_D \left(K(\mathbf{x}, \mathbf{x}) - \sum_{j=1}^{\infty} |e_j(\mathbf{x})|^2 \right) \varrho_D(d\mathbf{x}).$$

Taking Proposition 5.4 and (4.1) into account and noting that $\mathbb{P}(\|\mathbf{H}_m - \mathbf{I}_m\| \leq 1/2)$ is larger than $1 - \delta$, we obtain the assertion. \square

In addition to that we may easily get a deviation inequality by using Markov's inequality and standard arguments. It reads as follows.

Corollary 5.6. *Under the same assumptions as in Theorem 5.5 it holds for fixed $\delta > 0$*

$$\mathbb{P} \left(\sup_{\|f\|_{H(K)} \leq 1} \|f - S_{\mathbf{X}}^m f\|_{L_2(D, \varrho_D)}^2 \leq \frac{C}{\delta} \max \left\{ \sigma_m^2, \frac{\log(8n)}{n} T(m) \right\} \right) \geq 1 - 3\delta, \quad (5.9)$$

where $C := 3 + 8C_R^2 + 4C_R < 28.05$ is an absolute constant.

Proof. We define the events

$$A := \left\{ \mathbf{X} : \sup_{\|f\|_{H(K)} \leq 1} \|f - S_{\mathbf{X}}^m f\|_{L_2(D, \varrho_D)}^2 \leq t \right\},$$

$$B := \{\mathbf{X} : \|\mathbf{H}_m - \mathbf{I}_m\| \leq 1/2\}$$

and split up

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

Treating each summand separately, we have

$$\mathbb{P}(A^c \cap B) = \mathbb{P}(A^c | B) \mathbb{P}(B) \leq \mathbb{P}(A^c | B),$$

$$\mathbb{P}(A^c \cap B^c) \leq \mathbb{P}(B^c) \leq \delta.$$

Next we estimate $\mathbb{P}(A^c | B)$ using the Markov inequality (4.2), Theorem 5.5, and setting

$$t := \frac{3 + 8C_R^2 + 4C_R}{\delta} \max \left\{ \sigma_m^2, \frac{\log(8n)}{n} T(m) \right\}$$

which yields

$$\mathbb{P}(A^c | B) \leq \frac{\mathbb{E}(A^c | B)}{t} \leq \frac{\delta}{1 - \delta} \stackrel{\delta \leq 1/2}{\leq} 2\delta$$

and the assertion follows. \square

Example 5.7. Theorem 5.5 as well as Corollary 5.6 can also be applied to non-bounded orthonormal systems which may lead to non-optimal error bounds. For instance, let $D = [-1, 1]$ and ϱ_D the normalized Lebesgue measure on $D = [-1, 1]$. Then the second order operator A defined by

$$Af(x) = -((1 - x^2)v')'$$

characterizes for $s > 1$ weighted Sobolev spaces

$$H(K_s) := \{f \in L_2(D) : A^{s/2}f \in L_2(D)\}$$

which are in fact reproducing kernel Hilbert spaces with reproducing kernel

$$K_s(x, y) = \sum_{k \in \mathbb{N}} (1 + (k(k+1))^s)^{-1} \mathcal{P}_k(x) \mathcal{P}_k(y),$$

where $\mathcal{P}_k : D \rightarrow \mathbb{R}$, $k \in \mathbb{N}$, are $L_2(D)$ -normalized Legendre polynomials. Clearly, $(\mathcal{P}_k)_{k \in \mathbb{N}}$ provides an ONB in $L_2(D)$ and plays the role of $(\eta_k)_{k \in \mathbb{N}}$ in our setting. Moreover, we have $e_k = \lambda_k \eta_k = (1 + (k(k+1))^s)^{-1/2} \mathcal{P}_k$, and, accordingly, $\sigma_k = (1 + (k(k+1))^s)^{-1/2}$. The two quantities $N(m)$ and $T(m)$ are given by

$$N(m) = \sup_{x \in D} \sum_{k=1}^{m-1} |\mathcal{P}_k(x)|^2 = \sum_{k=0}^{m-2} \frac{2k+1}{2} = \frac{(m-1)^2}{2},$$

$$T(m) = \sup_{x \in D} \sum_{k=m}^{\infty} \frac{2k-1}{2(1 + (k(k+1))^s)} \asymp \sum_{k=m}^{\infty} k^{-2s+1} \asymp m^{-2s+2}.$$

Applying Theorem 5.5 or Corollary 5.6 and choosing m as large as possible, leads to the relation $m^2 \sim n/\log n$, which is far from optimal with respect to the number n of used samples, i.e., we observe worst case error estimates of the form

$$\sup_{\|f\|_{H(K_s)} \leq 1} \|f - S_{\mathbf{X}}^m f\|_{L_2(D)} \lesssim \sigma_m \sim m^{-s} \lesssim \left(\frac{\log(n)}{n} \right)^{s/2}$$

in expectation and with high probability, respectively.

On the other hand, the result by Bernardi and Maday [4, Thm. 6.2] using a polynomial interpolation operator j_{n-1} at Gauss points guarantees

$$\sup_{\|f\|_{H(K_s)} \leq 1} \|f - j_{n-1} f\|_{L_2(D)} \lesssim \frac{\log(n)}{n^s},$$

which is optimal in its main rate s . However, as it turns out below (see Example 5.10), it is not optimal with respect to the power of the logarithm. \square

Example 5.7 illustrates that the suggested approach will lead to worst case error estimates with high probability. However, the achieved upper bounds are not optimal in specific situations. We will overcome this limitation in the next section by drawing the sampling nodes with respect to a weighted, tailored distribution and apply a weighted least squares algorithm.

5.3 Improvements due to importance sampling

We are interested in the question of optimal sampling recovery of functions from reproducing kernel Hilbert spaces in $L_2(D, \varrho_D)$. The goal is to get reasonable bounds in n , preferably in terms of the singular numbers of the embedding. Theorem 5.5 already gives a satisfying answer in case of bounded kernels and $N(m) \in \mathcal{O}(m)$. In order to drop both conditions, we will use a weighted (deterministic) least squares algorithm (see Algorithm 2) to recover functions $f \in H(K)$ from samples at random nodes (“random information” in the sense of [27]). The approach is a slight modification of the one proposed earlier in [33]. A technique will be used, which is known as “(optimal) importance sampling”, where one defines a density function depending on the spectral properties of the embedding operator. The sampling nodes are then drawn according to this density. In the Monte-Carlo setting (or “randomized setting”) this has been successfully applied, e.g., by Cohen and Migliorati in [15], see Remark 6.3 below. Also in connection with compressed sensing it led to substantial improvements when recovering multivariate functions, see [56, 55]. Authors originally applied this technique e.g. for the approximation of integrals, see [26]. However, the setting in which we are interested in requires additional work since the sampling nodes are supposed to be drawn in advance for the whole class of functions.

As already mentioned, we construct a more suitable distribution which is used to draw the sampling nodes at random. In particular, we tailor a probability density function $\varrho_m: D \rightarrow \mathbb{C}$ such that μ_m , which is given by

$$\mu_m(A) := \int_A \varrho_m(\mathbf{x}) \varrho_D(d\mathbf{x}) \quad , \quad A \subset D \text{ measurable} . \quad (5.11)$$

Then we may draw the sampling nodes in $\mathbf{X} \subset D$ i.i.d. at random according to μ_m . For the chosen set \mathbf{X} , we define the approximation operator $\tilde{S}_{\mathbf{X}}^m$ as indicated in Algorithm 2.

Algorithm 2 Weighted least squares regression.

Input: $\mathbf{X} = \{\mathbf{x}^1, \dots, \mathbf{x}^n\} \subset D$ set of distinct sampling nodes,
 $\mathbf{f} = (f(\mathbf{x}^1), \dots, f(\mathbf{x}^n))^\top$ 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 (5.10) has full
(column) rank.

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

Solve the over-determined linear system

$$\tilde{\mathbf{L}}_m (\tilde{c}_1, \dots, \tilde{c}_{m-1})^\top = \mathbf{g}, \quad \tilde{\mathbf{L}}_m := (l_{k,j})_{k=1, j=1}^{n, m-1}, \quad l_{k,j} := \begin{cases} 0, & \varrho_m(\mathbf{x}^k) = 0, \\ \eta_j(\mathbf{x}^k)/\sqrt{\varrho_m(\mathbf{x}^k)}, & \varrho_m(\mathbf{x}^k) \neq 0, \end{cases} \quad (5.10)$$

via least squares (e.g. directly or via the LSQR algorithm [52]), i.e., compute

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

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

Choosing the specific density function

$$\varrho_m(\mathbf{x}) := \frac{1}{2} \left(\frac{1}{m-1} \sum_{j=1}^{m-1} |\eta_j(\mathbf{x})|^2 + \left(\sum_{j=m}^{\infty} \lambda_j \right)^{-1} \left(K(\mathbf{x}, \mathbf{x}) - \sum_{j=1}^{m-1} |e_j(\mathbf{x})|^2 \right) \right) \quad (5.12)$$

guarantees worst case error estimates which are optimal up to logarithmic factors and up to a specific failure probability.

Theorem 5.8. *Let $H(K)$ be a separable reproducing kernel Hilbert space of complex-valued functions defined on D such that*

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

for some non-trivial σ -finite measure ϱ_D on D , where $(\sigma_j)_{j \in \mathbb{N}}$ denotes the non-increasing sequence of singular numbers of the embedding $\text{Id} : H(K) \rightarrow L_2(D, \varrho_D)$. Let further $\delta \in (0, 1/3)$ and $n \in \mathbb{N}$ large enough, such that

$$m := \left\lfloor \frac{n}{96(\sqrt{2} \log(2n) - \log \delta)} \right\rfloor \geq 2 \quad (5.13)$$

holds. Moreover, we assume $\varrho_m : D \rightarrow \mathbb{C}$ and μ_m as stated in (5.12) and (5.11). We draw each node in $\mathbf{X} := \{\mathbf{x}^1, \dots, \mathbf{x}^n\} \subset D$ i.i.d. at random according to μ_m , which yields

$$\mathbb{P} \left(\sup_{\|f\|_{H(K)} \leq 1} \|f - \tilde{S}_{\mathbf{X}}^m f\|_{L_2(D, \varrho_D)}^2 \leq \frac{C}{\delta} \max \left\{ \sigma_m^2, \frac{\log(8n)}{n} \sum_{j=m}^{\infty} \sigma_j^2 \right\} \right) \geq 1 - 3\delta, \quad (5.14)$$

where $C := 3 + 16 C_R^2 + 4\sqrt{2} C_R < 49.5$ is an absolute constant.

Proof. We emphasize that the second argument of the max term in (5.14) makes sense since we know from (2.6) that the sequence of singular numbers is square-summable. As a technical modification of the density function, which has been presented in [33], we use the density $\varrho_m : D \rightarrow \mathbb{R}$ as stated in (5.12). 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 $W_{\varrho_D} := \text{Id}^* \circ \text{Id} : H(K) \rightarrow H(K)$, the sequence $(\lambda_j)_{j \in \mathbb{N}}$ represents the ordered eigenvalues, and finally $\eta_j := \lambda_j^{-1/2} e_j$. Since we assume the separability of $H(K)$ and the σ -finiteness of ϱ_D we observe equality in (2.6), cf. [11, Thm. 4.27], and thus we easily see that $\int_D \varrho_m(\mathbf{x}) \varrho_D(d\mathbf{x}) = 1$. Let us define a family of kernels $\tilde{K}_m(\mathbf{x}, \mathbf{y})$, indexed by $m \in \mathbb{N}$, via

$$\tilde{K}_m(\mathbf{x}, \mathbf{y}) := \frac{K(\mathbf{x}, \mathbf{y})}{\sqrt{\varrho_m(\mathbf{x}) \varrho_m(\mathbf{y})}}, \quad (5.15)$$

and a new measure μ_m as stated in (5.11) with the corresponding weighted space $L_2(D, \mu_m)$. Clearly, $\tilde{K}_m(\cdot, \cdot)$ is a positive type function. As a consequence of

$$|K(\mathbf{x}, \mathbf{y})| \leq \sqrt{K(\mathbf{x}, \mathbf{x})} \cdot \sqrt{K(\mathbf{y}, \mathbf{y})}, \quad \mathbf{x}, \mathbf{y} \in D,$$

we obtain by an elementary calculation that with a constant $c_m > 0$ it holds

$$|K(\mathbf{x}, \mathbf{y})| \leq c_m \sqrt{\varrho_m(\mathbf{x})} \cdot \sqrt{\varrho_m(\mathbf{y})}. \quad (5.16)$$

Indeed,

$$\begin{aligned} K(\mathbf{x}, \mathbf{x}) &= \sum_{j=1}^{m-1} |e_j(\mathbf{x})|^2 + \left(K(\mathbf{x}, \mathbf{x}) - \sum_{j=1}^{m-1} |e_j(\mathbf{x})|^2 \right) \\ &= \sum_{j=1}^{m-1} \lambda_j |\eta_j(\mathbf{x})|^2 + \left(\sum_{j=m}^{\infty} \lambda_j \right) \cdot \left(\sum_{j=m}^{\infty} \lambda_j \right)^{-1} \left(K(\mathbf{x}, \mathbf{x}) - \sum_{j=1}^{m-1} |e_j(\mathbf{x})|^2 \right) \\ &\leq c_m \varrho_m(\mathbf{x}) \end{aligned}$$

with

$$c_m := 2 \max \left\{ \lambda_1(m-1), \sum_{j=m}^{\infty} \lambda_j \right\}.$$

Hence, if $\varrho_m(\mathbf{x})$ or $\varrho_m(\mathbf{y})$ happens to be zero then we may put $\tilde{K}_m(\mathbf{x}, \mathbf{y}) := 0$ in (5.15). In any case, due to (5.16), the kernel $\tilde{K}_m(\mathbf{x}, \mathbf{y})$ fits the requirements in Theorem 5.5. In fact, in Theorem 5.5 it is necessary that $\tilde{N}(m)$ and $\tilde{T}(m)$ are well-defined and that we have access to function values in order to create the matrices $\tilde{\mathbf{L}}_m$ and take the function values $f(\mathbf{x}^k)$, $k = 1, \dots, n$. Let us discuss the quantities $\tilde{N}(m)$ and $\tilde{T}(m)$ appearing in Theorem 5.5 for this new kernel $\tilde{K}_m(\mathbf{x}, \mathbf{y})$ first. It is clear, that the embedding $\text{Id} : H(\tilde{K}_m) \rightarrow L_2(D, \mu_m)$ shares the same spectral properties as the original embedding. Note that a function g belongs to $H(\tilde{K}_m)$ if and only if $g(\cdot) = f(\cdot)/\sqrt{\varrho_m(\cdot)}$, $f \in H(K)$, where we always put $0/0 := 0$. Clearly, as a consequence of (5.16) and (5.17) below (together with a density argument), we have that $\varrho(\mathbf{x}) = 0$ implies $f(\mathbf{x}) = 0$ for all $f \in H(K)$. Moreover, whenever $\|f\|_{H(K)} \leq 1$, the function $g := f/\sqrt{\varrho_m}$ satisfies $\|g\|_{H(\tilde{K}_m)} \leq 1$. Indeed, let

$$f(\cdot) = \sum_{i=1}^N \alpha_i K(\cdot, \mathbf{x}^i). \quad (5.17)$$

Then $\langle f, f \rangle_{H(K)} = \sum_{j=1}^N \sum_{i=1}^N \alpha_i \overline{\alpha_j} K(\mathbf{x}^j, \mathbf{x}^i)$. We have

$$\begin{aligned} g &= f(\cdot) / \sqrt{\varrho_m(\cdot)} = \sum_{i=1}^N \alpha_i K(\cdot, \mathbf{x}^i) / \sqrt{\varrho_m(\cdot)} \\ &= \sum_{i=1}^N \alpha_i \sqrt{\varrho_m(\mathbf{x}^i)} \frac{K(\cdot, \mathbf{x}^i)}{\sqrt{\varrho_m(\cdot)} \sqrt{\varrho_m(\mathbf{x}^i)}}. \end{aligned}$$

This implies

$$\begin{aligned} \langle g, g \rangle_{H(\tilde{K}_m)} &= \sum_{j=1}^N \sum_{i=1}^N \alpha_i \overline{\alpha_j} \sqrt{\varrho_m(\mathbf{x}^i)} \sqrt{\varrho_m(\mathbf{x}^j)} \frac{K(\mathbf{x}^j, \mathbf{x}^i)}{\sqrt{\varrho_m(\mathbf{x}^j)} \sqrt{\varrho_m(\mathbf{x}^i)}} \\ &= \langle f, f \rangle_{H(K)}. \end{aligned}$$

What remains is a standard density argument. The singular numbers of the new embedding remain the same. The singular vectors $\tilde{e}_k(\cdot)$ and $\tilde{\eta}_k(\cdot)$ are slightly different. They are the original ones divided by $\sqrt{\varrho_m(\cdot)}$. In fact,

$$\tilde{N}(m) := \sup_{\mathbf{x} \in D} \sum_{k=1}^{m-1} |\eta_k(\mathbf{x})|^2 / \varrho_m(\mathbf{x}) \leq 2(m-1) \sup_{\mathbf{x} \in D} \sum_{k=1}^{m-1} |\eta_k(\mathbf{x})|^2 / \sum_{j=1}^{m-1} |\eta_j(\mathbf{x})|^2 = 2(m-1).$$

Furthermore, taking (2.5) into account, we find

$$\begin{aligned} \tilde{T}(m) &:= \sup_{\mathbf{x} \in D} \sum_{k=m}^{\infty} |e_k(\mathbf{x})|^2 / \varrho_m(\mathbf{x}) \\ &\leq 2 \left(\sum_{j=m}^{\infty} \lambda_j \right) \sup_{\mathbf{x} \in D} \sum_{k=m}^{\infty} |e_k(\mathbf{x})|^2 / \left(K(\mathbf{x}, \mathbf{x}) - \sum_{j=1}^{m-1} |e_j(\mathbf{x})|^2 \right) \\ &\leq 2 \left(\sum_{j=m}^{\infty} \lambda_j \right) \sup_{\mathbf{x} \in D} \sum_{k=m}^{\infty} |e_k(\mathbf{x})|^2 / \sum_{j=m}^{\infty} |e_j(\mathbf{x})|^2 \\ &\leq 2 \sum_{j=m}^{\infty} \lambda_j = 2 \sum_{j=m}^{\infty} \sigma_j^2. \end{aligned} \tag{5.18}$$

In order to define the new reconstruction operator $\tilde{S}_{\mathbf{X}}^m$ we need to create the matrices $\tilde{\mathbf{L}}_m$ using the new function system $\tilde{\eta}_k$ and take function evaluations $f(\mathbf{x}^1), \dots, f(\mathbf{x}^n)$. In more detail, we solve the least squares problem

$$\tilde{\mathbf{L}}_m \tilde{\mathbf{c}} = \mathbf{g}, \text{ where } \tilde{\mathbf{L}}_m := \left(\tilde{\eta}_j(\mathbf{x}^k) \right)_{k=1, j=1}^{n, m-1}, \quad \mathbf{g} := \left(\frac{f(\mathbf{x}^1)}{\sqrt{\varrho_m(\mathbf{x}^1)}}, \dots, \frac{f(\mathbf{x}^n)}{\sqrt{\varrho_m(\mathbf{x}^n)}} \right)^{\top},$$

and the vector $\tilde{\mathbf{c}}$ contains the coefficients of the least squares approximation $S_{\mathbf{X}}^m(g) = \sum_{j=1}^{m-1} \tilde{c}_j \tilde{\eta}_j$ of $g := f / \sqrt{\varrho_m}$. This leads to Algorithm 2. Consequently, Theorem 5.5 allows to estimate the error

$$\|g - S_{\mathbf{X}}^m(g)\|_{L_2(D, \mu_m)} = \|f - \sqrt{\varrho_m} S_{\mathbf{X}}^m(g)\|_{L_2(D, \varrho_D)} = \|f - \tilde{S}_{\mathbf{X}}^m f\|_{L_2(D, \varrho_D)},$$

where $\tilde{S}_{\mathbf{X}}^m f := \sqrt{\varrho_m} S_{\mathbf{X}}^m(g) = \sum_{j=1}^{m-1} \tilde{c}_j \tilde{\eta}_j(\mathbf{x})$.

We stress that $\tilde{S}_{\mathbf{X}}^m f$ and the direct computation of $S_{\mathbf{X}}^m f$ using $\mathbf{L}_m \mathbf{c} = \mathbf{f}$ may not coincide since both are based on different least squares problems in general.

It remains to note that for fixed n and m as in (5.13) we have for $\mathbf{X} = (\mathbf{x}^1, \dots, \mathbf{x}^n)$ the relation

$$\begin{aligned} \sup_{\|f\|_{H(K)} \leq 1} \|f - \tilde{S}_{\mathbf{X}}^m(f)\|_{L_2(D, \varrho_D)}^2 &= \sup_{\|f\|_{H(K)} \leq 1} \|f/\sqrt{\varrho_m} - S_{\mathbf{X}}^m(f/\sqrt{\varrho_m})\|_{L_2(D, \mu_m)}^2 \\ &\leq \sup_{\|g\|_{H(\tilde{K}_m)} \leq 1} \|g - S_{\mathbf{X}}^m(g)\|_{L_2(D, \mu_m)}^2. \end{aligned}$$

Applying a slight modification of Corollary 5.6, i.e., setting

$$t := \frac{1}{\delta} \left(3\sigma_m^2 + 8C_R^2 \frac{\log(8n)}{n} \tilde{T}(m) + 4C_R \sigma_m \sqrt{\frac{\log(8n)}{n} \tilde{T}(m)} \right)$$

in the proof, yields

$$\mathbb{P} \left(\sup_{\|f\|_{H(K)} \leq 1} \|f - \tilde{S}_{\mathbf{X}}^m f\|_{L_2(D, \varrho_D)}^2 \leq \frac{C}{\delta} \max \left(\sigma_m^2, \frac{\log(8n)}{n} \sum_{j=m}^{\infty} \sigma_j^2 \right) \right) \geq 1 - 3\delta,$$

where we took (5.18) into account and $C = 3 + 16C_R^2 + 4\sqrt{2}C_R$ is as stated above. \square

Remark 5.9. (i) In order to prove the theorem in full generality for separable RKHS, we use a technical modification of the density function presented in [33]. Clearly, as a consequence of (2.5) 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. The pointwise defined density function will be an essential ingredient for drawing the nodes in \mathbf{X} on the one hand and performing a reweighted least squares fit on the other hand. Note that also point evaluations of the density function are used in the algorithm. To circumvent the lacking injectivity we introduce a new reproducing kernel Hilbert space $H(\tilde{K}_m)$ built upon the modified density function. To this situation Corollary 5.6 is applied, and we obtain Theorem 5.8, which also improves the results in [33] by determining explicit constants. In turn, we show, that the original algorithm proposed by [33] also works in this more general situation since both densities are equal almost everywhere.

(ii) The situation of a non-separable RKHS $H(K)$ is not considered in Theorem 5.8. It has been considered in the subsequent paper [43] by Moeller and T. Ullrich. Here the sampling density has to be modified even further to get a bound

$$\sup_{\|f\|_{H(K)} \leq 1} \|f - \tilde{S}_{\mathbf{X}}^m f\|_{L_2(D, \varrho_D)}^2 \leq c_1 \max \left\{ \frac{1}{n}, \frac{1}{m} \sum_{k \geq m/2} \sigma_k^2 \right\}$$

with probability larger than $1 - c_2 n^{1-r}$ and $m := \lfloor n/(c_3 r \log(n)) \rfloor$. With this method we can not get beyond the rate $n^{-1/2}$ in case of non-separable Hilbert spaces $H(K)$.

(iii) While this paper was under review, the statement ‘‘recovery with high probability’’ has been refined by M. Ullrich [64] and, independently, by Moeller, T. Ullrich [43]. In fact, the failure probability of the above recovery can be controlled by n^{-r} (and is therefore decaying polynomially in the number of samples) by only paying a multiplicative constant r in the bounds, see also (ii).

(iv) As a further follow up, Nagel, Schäfer and T. Ullrich [44] developed a subsampling technique to select a subset $\mathbf{J} \subset \mathbf{X}$ of $\mathcal{O}(m)$ nodes out of the randomly chosen node set \mathbf{X} such that the corresponding least squares recovery operator $\tilde{S}_{\mathbf{J}}^m$ gives

$$\sup_{\|f\|_{H(K)} \leq 1} \|f - \tilde{S}_{\mathbf{J}}^m f\|_{L_2(D, \varrho_D)}^2 \leq \frac{C \log m}{m} \sum_{k=cm}^{\infty} \sigma_k^2$$

with precisely determined universal constants $C, c > 0$.

Example 5.10. (i) If we choose $m \asymp n/\log(n)$ according to (5.13) and assume a polynomial decay of the singular values, i.e., $\sigma_m \lesssim m^{-s} \log^\alpha m$ with $s > 1/2$ (which is for instance the case for Sobolev embeddings into L_2), we find

$$\sup_{\|f\|_{H(K_s)} \leq 1} \|f - \tilde{S}_{\mathbf{X}}^m f\|_{L_2(D)} \lesssim m^{-s} \log^\alpha m \asymp n^{-s} \log^{\alpha+s} n.$$

Accordingly, we observe the best possible main rate $-s$ with respect to the used number of samples n , i.e., we achieve optimality up to logarithmic factors. For a more specific example, see (ii) below and Section 8.

(ii) Let us reconsider the setting from Example 5.7. Theorem 5.8 provides a highly improved sampling strategy compared to the one discussed in Example 5.7. Certainly, we change the underlying distribution for the random selection of the sampling nodes and we incorporate weights in the least squares algorithm, cf. Algorithm 2. However, this allows for a crucial improvement of the relation of the maximal polynomial degree m and the number n of sampling nodes to $m \sim n/\log(n)$, which leads to estimates

$$\sup_{\|f\|_{H(K_s)} \leq 1} \|f - \tilde{S}_{\mathbf{X}}^m f\|_{L_2(D)} \lesssim \max \left\{ \sigma_m, \sqrt{m^{-1} \sum_{k=m}^{\infty} \sigma_k^2} \right\} \asymp m^{-s} \lesssim \left(\frac{\log(n)}{n} \right)^s$$

that are optimal in m and – up to logarithmic factors – even optimal in n . Here $s > 1/2$ is admitted. Compared to the results in Example 5.7, we obtain the same rates of convergence with respect to the polynomial degree m and significantly improved rates of convergence with respect to the number n of used sampling values. Note, that instead of the density given in (5.12) we may use the Chebychev measure given by the density $\varrho(\mathbf{x}) = (\pi\sqrt{1-x^2})^{-1}$ on $D = [-1, 1]$ since the L_2 -normalized Legendre polynomials $\mathcal{P}_k(x)$ are dominated by $C(1-x^2)^{-1/4}$ for all $k \in \mathbb{N}_0$, see [55]. Hence, in contrast to (5.12), this sampling measure is universal for all m .

In comparison to the results in Example 5.7, we obtain a significantly better rate of convergence with respect to the number n of used sampling values. Applying the recent result mentioned in Remark 5.9(iv) we obtain for the subsampled weighted least squares operator the worst-case error bound

$$\sup_{\|f\|_{H(K_s)} \leq 1} \|f - \tilde{S}_{\mathbf{J}}^m f\|_{L_2(D)} \lesssim \frac{\sqrt{\log(n)}}{n^s}.$$

Note, that $\tilde{S}_{\mathbf{J}}^m f$ uses $n \in \mathcal{O}(m)$ many samples of f . □

5.4 The power of standard information and tractability

The approximation framework studied in this paper has been first considered by Wasilkowski, Woźniakowski in 2001, see [65]. After that many authors dealt with the problem on how well can we approximate functions from RKHS by only using function values. For further historical remarks see [49] and [44, Rem. 1].

In this context, sampling values are called *standard information* abbreviated by Λ^{std} . In contrast to this, one may also allow linear information (abbreviated by Λ^{all}), which means that the approximation is computed from a number of information about the function coming from arbitrary linear functionals. Clearly, Λ^{std} is a subset of Λ^{all} due to (2.1). It is well known that the best possible worst case error with respect to m information coming from Λ^{all} can be achieved by approximating functions by means of their corresponding exact Fourier partial sum within $\text{span}\{\eta_j : j = 1, \dots, m\}$. Then the corresponding worst case error is determined by σ_{m+1} , see [46, Cor. 4.12], which establishes a natural lower bound on the worst case error with respect to Λ^{std} . One crucial question in IBC is to determine whether or not standard information Λ^{std} is as powerful as linear information Λ^{all} . In this context, “power” is usually identified with the order of convergence of the considered error with respect to the number of used information.

From Theorem 5.8, see also Example 5.10,(i), we obtain that the polynomial order of convergence for standard information is the same as for linear information if the kernel has a finite trace (i.e., $\text{Id} : H(K) \rightarrow L_2(D, \varrho_D)$ is a Hilbert-Schmidt operator). This problem has been addressed in [37], [48, Open Problem 1] and [49, Open Problem 126]. In [33] this observation has been already made for the situation that the embedding operator is injective (such that the eigenvectors of the strictly positive eigenvalues form an orthonormal basis in $H(K)$, see Remark 2.1 above). The contribution of Theorem 5.8 is to get explicitly determined constants on the one hand. On the other hand, it shows that separability of the RKHS and a finite trace condition is essentially enough for this purpose. Note that the finite trace condition can not be avoided, see [28].

We will further discuss some consequences for the tractability of this problem. For the necessary notions and definitions from the field of Information Based Complexity, see [46, 47, 49]. We comment on polynomial tractability with respect to linear information Λ^{all} and standard information Λ^{std} . Let us consider the family of approximation problems

$$\text{APP}_d : H(K_d) \rightarrow L_2(D_d, \varrho_{D_d}), \quad d \in \mathbb{N},$$

where $K_d : D_d \times D_d \rightarrow \mathbb{C}$ is a family of reproducing kernels. In [46, Thm. 5.1] *strong polynomial tractability* of the family $\{\text{APP}_d\}$ with respect to Λ^{all} is characterized as follows: There is a $\tau > 0$ such that

$$C := \sup_d \left(\sum_{j=1}^{\infty} \sigma_{j,d}^{\tau} \right)^{1/\tau} < \infty, \quad (5.19)$$

where $\sigma_{j,d}$, $j = 1, \dots$, are the singular values belonging to APP_d for fixed d . From our analysis in Theorem 5.8 above we directly obtain a sufficient condition for polynomial tractability with respect to Λ^{std} . Without going into detail, we denote by $n^{\text{wor}}(\varepsilon, d; \Lambda)$, $\Lambda \in \{\Lambda^{\text{std}}, \Lambda^{\text{all}}\}$, the minimal number n of information out of the specified class Λ any algorithm requires in order to achieve a worst case error ε for the d -dimensional approximation problem. Certainly, $n^{\text{wor}}(\varepsilon, d; \Lambda)$ usually depends on ε , the dimension d and Λ . At this point, we would like to mention that polynomial tractability means that $n^{\text{wor}}(\varepsilon, d; \Lambda) \leq C \varepsilon^{-p} d^q$, $C, p, q \geq 0$, i.e.,

$n^{\text{wor}}(\varepsilon, d; \Lambda)$ can be bounded by terms that are both polynomial in d and polynomial in ε . Furthermore, the problem APP_d is called strongly polynomial tractable in the case that the estimate on n^{wor} holds with $q = 0$.

Theorem 5.11. *The family $\{\text{APP}_d\}$ is strongly polynomially tractable with respect to Λ^{std}*

- (i) *if there exists $\tau \leq 2$ such that (5.19) holds true or*
- (ii) *if $\{\text{APP}_d\}$ is strongly polynomially tractable with respect to Λ^{all} with exponent $0 < p_{\text{all}} < 2$, i.e., $n^{\text{wor}}(\varepsilon, d; \Lambda^{\text{all}}) \leq C_{\text{all}} \varepsilon^{-p_{\text{all}}}$.*

More precisely, there are constants $C_{\text{std}}, \delta > 0$ only depending on (C, τ) or $(C_{\text{all}}, p_{\text{all}})$ such that

$$n^{\text{wor}}(\varepsilon, d; \Lambda^{\text{std}}) \leq C_{\text{std}} \varepsilon^{-p_{\text{std}}} \log(1/\varepsilon)^\delta$$

with $p_{\text{std}} = \tau$ in case (i) and $p_{\text{std}} = p_{\text{all}}$ in case (ii).

Proof. Note that (ii) implies (i) with $\tau = 2$. Furthermore, Theorem 5.8 implies strong polynomial tractability if (i) is assumed. In case (i) we may use Stechkin's lemma [17, Lem. 7.8] which gives that $\sum_{j=m}^{\infty} \sigma_{j,d}^2 \leq C^2 m^{-2/\tau+1}$ for all d . This gives the exponent $p_{\text{std}} = \tau$ and an additional log due to (5.13). If (ii) is assumed then $\sum_{j=m}^{\infty} \sigma_{j,d}^2 \leq C' m^{-2/p_{\text{all}}+1}$ for all d . \square

Theorem 5.11 is stronger than [49, Thm. 26.20] in two aspects. As pointed out in the proof, assumption (i) is weaker than (ii), which is essentially the one in [49, Thm. 26.20]. Furthermore, our statement is stronger since p_{std} equals p_{all} . The authors in [49, 26.6.1] showed that $p_{\text{std}} = p_{\text{all}} + \frac{1}{2}[p_{\text{all}}]^2$ and proposed that “the lack of the exact exponent represents another major challenge” and formulated Open Problem 127. Our considerations prove that the dependence on ε of the tractability estimates on $n^{\text{wor}}(\varepsilon, d; \Lambda^{\text{all}})$ and $n^{\text{wor}}(\varepsilon, d; \Lambda^{\text{std}})$ coincide up to logarithmic factors. Similar assertions hold true when *strong polynomial tractability* is replaced by *polynomial tractability*. The modifications are straight forward.

Example 5.12. Let us consider an example from [53] and [36], namely, if $s_1 \leq s_2 \leq \dots \leq s_d$ then

$$H(K_d) := H_{\text{mix}}^{\vec{s}}(\mathbb{T}^d) = H^{s_1}(\mathbb{T}) \otimes \dots \otimes H^{s_d}(\mathbb{T})$$

and $\text{APP}_d : H_{\text{mix}}^{\vec{s}}(\mathbb{T}^d) \rightarrow L_2(\mathbb{T}^d)$. Here $\mathbb{T} = [0, 1]$, where opposite points are identified, and $H^s(\mathbb{T})$ is normed by

$$\|f\|_{H^s}^2 := \sum_{k \in \mathbb{Z}} |\hat{f}_k|^2 w_s(k)^2$$

with $w_s(k) = \max\{1, (2\pi)^s |k|^s\}$, see also Section 8. The smoothness vector \vec{s} is supposed to grow in j , e.g., for $\beta > 0$ we have

$$s_j \geq \beta \log_2(j+1), \quad j \in \mathbb{N}. \quad (5.20)$$

It has been shown in [53], see also [36], that the growth condition (5.20) is necessary and sufficient for polynomial tractability with respect to Λ^{all} . Taking Theorem 5.11 into account we easily check that (5.19) is satisfied with $\tau \leq 2$ whenever $\beta \cdot \tau > 1$, which means that $\beta > 1/2$ is sufficient for strong polynomial tractability. In this case we obtain for any $1/\beta < \tau \leq 2$ that

$$n^{\text{wor}}(\varepsilon, d; \Lambda^{\text{std}}) \leq C_\tau \varepsilon^{-\tau}.$$

6 Recovery of individual functions

In this section we are interested in the reconstruction of an individual function f (taken from the unit ball of $H(K)$) from samples at random nodes. In the IBC community, see [46, 47, 49], such a scenario is called “randomized setting”, which refers to the occurrence of a random element in the algorithm (Monte Carlo). The following investigations improve on the results in [66] from different points of view. On the one hand, it is not necessary to choose sampling nodes according to a whole bunch of probability density functions according to Remark 6.3. On the other hand, assuming a bounded kernel turns out to be sufficient to obtain the rate of convergence matching that of $(\sigma_k)_{k \in \mathbb{N}}$ when randomly choosing sampling nodes according to the given probability measure ϱ_D due to Corollary 6.2.

However, the subsequent analysis is related to the one in [13, 14]. With similar techniques as above we will get an estimate for the conditional mean of the individual error $\|f - S_{\mathbf{X}}^m f\|_{L_2(D, \varrho_D)}$.

Theorem 6.1. *Let ϱ_D be a probability measure on D and $H(K)$ denote a reproducing kernel Hilbert which is compactly embedded into the space $L_2(D, \varrho_D)$. Let $m, n \in \mathbb{N}$, $m \geq 2$, be chosen such that (5.1) holds for some $0 < \delta < 1$. Let further f be a fixed function such that $\|f\|_{H(K)} \leq 1$. Drawing each sampling node in $\mathbf{X} = \{\mathbf{x}^1, \dots, \mathbf{x}^n\} \subset D$ i.i.d. at random according to ϱ_D , we have for the conditional expectation of the individual error*

$$\mathbb{E}\left(\|f - S_{\mathbf{X}}^m f\|_{L_2(D, \varrho_D)}^2 \middle| \|\mathbf{H}_m - \mathbf{I}_m\| \leq 1/2\right) \leq \frac{1}{1-\delta} \left(\sigma_m^2 + \frac{C_\delta}{\log n} \sigma_m^2 \right) \leq \frac{1.1}{1-\delta} \sigma_m^2,$$

where $0 < C_\delta \leq 0.06$ depends on δ .

Proof. This time we follow the proof of [13, Thm. 2] and obtain

$$\begin{aligned} \|f - S_{\mathbf{X}}^m f\|_{L_2(D, \varrho_D)}^2 &= \|f - P_{m-1} f\|_{L_2(D, \varrho_D)}^2 + \|P_{m-1} f - S_{\mathbf{X}}^m f\|_{L_2(D, \varrho_D)}^2 \\ &\leq \sigma_m^2 + \|S_{\mathbf{X}}^m(P_{m-1} f - f)\|_{L_2(D, \varrho_D)}^2 \\ &\leq \sigma_m^2 + \|(\mathbf{L}_m^* \mathbf{L}_m)^{-1}\|^2 \|\mathbf{L}_m^* (g(\mathbf{x}^1), \dots, g(\mathbf{x}^n))^{\top}\|_2^2 \\ &\leq \sigma_m^2 + \frac{4}{n^2} \sum_{k=1}^{m-1} \left| \sum_{j=1}^n \overline{\eta_k(\mathbf{x}^j)} g(\mathbf{x}^j) \right|^2 \\ &= \sigma_m^2 + \frac{4}{n^2} \sum_{k=1}^{m-1} \sum_{j=1}^n \sum_{i=1}^n g(\mathbf{x}^i) \overline{g(\mathbf{x}^j) \eta_k(\mathbf{x}^i)} \eta_k(\mathbf{x}^j), \end{aligned}$$

where $g = f - P_{m-1} f$. Averaging on both sides yields

$$\begin{aligned} &\int_{\|\mathbf{H}_m - \mathbf{I}_m\| \leq 1/2} \|f - S_{\mathbf{X}}^m f\|_{L_2(D, \varrho_D)}^2 \varrho_D^n(d\mathbf{x}) \\ &\leq \sigma_m^2 + \frac{4}{n^2} \sum_{k=1}^{m-1} \sum_{j=1}^n \sum_{i=1}^n \mathbb{E}(g(\mathbf{x}^i) \overline{g(\mathbf{x}^j) \eta_k(\mathbf{x}^i)} \eta_k(\mathbf{x}^j)) \\ &= \sigma_m^2 + \frac{4}{n} \sum_{k=1}^{m-1} \int_D |g(\mathbf{x})|^2 |\eta_k(\mathbf{x})|^2 \varrho_D(d\mathbf{x}) + \frac{4n(n-1)}{n^2} \sum_{k=1}^{m-1} \left| \int_D g(\mathbf{x}) \overline{\eta_k(\mathbf{x})} \varrho_D(d\mathbf{x}) \right|^2. \end{aligned}$$

Note that the last summand on the right-hand side vanishes since g is orthogonal to η_k , $k = 1, \dots, m-1$. This gives

$$\begin{aligned} & \int_{\|\mathbf{H}_m - \mathbf{I}_m\| \leq 1/2} \|f - S_{\mathbf{X}}^m f\|_{L_2(D, \varrho_D)}^2 \varrho_D^n(d\mathbf{x}) \\ & \leq \sigma_m^2 + \frac{4N(m)}{n} \|f - P_{m-1} f\|_{L_2(D, \varrho_D)}^2 \\ & \leq \sigma_m^2 + \frac{C_\delta}{\log(2n)} \sigma_m^2 \end{aligned}$$

by taking $4N(m)/n \leq C_\delta/\log(2n)$ into account, see (5.1), and there exists a $C_\delta \leq \frac{4}{48\sqrt{2}} \leq 0.06$. \square

Analogously to the proof of Corollary 5.6, one shows the following result.

Corollary 6.2. *Under the same assumptions as in Theorem 6.1 it holds for fixed $\delta > 0$*

$$\mathbb{P}\left(\|f - S_{\mathbf{X}}^m f\|_{L_2(D, \varrho_D)}^2 \leq \frac{1}{\delta} \sigma_m^2 \left(1 + \frac{0.06}{\log n}\right)\right) \geq 1 - 3\delta.$$

Remark 6.3. (i) Following [15] we may relax the condition (5.1) to

$$m := \left\lfloor \frac{n}{48(\sqrt{2}\log(2n) - \log \delta)} \right\rfloor + 1,$$

when sampling with respect to the new measure (importance sampling)

$$\mu_m(A) := \int_A \varrho_m(\mathbf{x}) \varrho_D(d\mathbf{x}),$$

where ϱ_m is given by

$$\varrho_m(\mathbf{x}) := \frac{1}{m-1} \sum_{k=1}^{m-1} |\eta_k(\mathbf{x})|^2.$$

Then, the corresponding approximation of the function f is based on the solution of a weighted least squares problem similar to the one discussed in Section 5.3. Note that we do not have to assume the square summability of the singular numbers of the embedding $\text{Id} : H(K) \rightarrow L_2(K, \varrho_D)$ here.

(ii) The result of Theorem 6.1 advanced with (i) directly leads to estimates on the power of standard information, see Section 5.4, in the so-called randomized setting, cf. [49]. Roughly speaking, n sampling values (standard information) in the randomized setting are at least as powerful as $cn/\log n$ linear information, i.e., the supremum over all $f \in H(K)$ of the expected approximation error that is caused by the weighted least squares regression is almost as good as the best possible worst case approximation error caused by the projection. The recent work [39] treats that topic in full detail. Moreover, in [38] the authors present improvements on the power of standard information in the so-called average case setting. Both paper investigate the weighted least squares regression given in Algorithm 2 using the techniques that yielded the results of this section.

(iii) Together with the Weaver subsampling technique in [44] we obtain a further significant improvement of the results in [39] in the situation above if the sequence $(\sigma_k)_{k \in \mathbb{N}}$ decays

“faster” than $k^{-1/2}$. Just for the sake of completeness, we state the corresponding general relation

$$e^{\text{ran}}(n, d; \Lambda^{\text{std}}) \leq C \min\{\sqrt{\log n} \cdot e^{\det}(c_1 n, d; \Lambda^{\text{all}}), e^{\det}(c_2 n / \log n, d; \Lambda^{\text{all}})\},$$

with three universal constants $C, c_1, c_2 > 0$, where we use the notation from [46, 47, 49].

7 Optimal weights for numerical integration

We consider the problem of approximating the integral with respect to μ_D of a function f by the cubature rule $Q_{\mathbf{X}}^m$

$$\text{Int}_{\mu_D} f := \int_D f \, d\mu_D \approx Q_{\mathbf{X}}^m f := \sum_{j=1}^n q_j f(\mathbf{x}^j) = \mathbf{q}^\top \mathbf{f},$$

where the weights q_j are determined by $\mathbf{X} = \{\mathbf{x}^1, \dots, \mathbf{x}^n\}$. Indeed, assuming \mathbf{L}_m of full column rank and following Oettershagen [50], we set

$$\begin{aligned} \mathbf{q} &:= \overline{\mathbf{L}_m} (\overline{\mathbf{L}_m^* \mathbf{L}_m})^{-1} \mathbf{b} \in \mathbb{C}^n, \\ \text{where } \mathbf{b} &:= (b_k)_{k=1}^{m-1} \in \mathbb{C}^{m-1} \text{ with } b_k := \int_D \eta_k \, d\mu_D. \end{aligned} \quad (7.1)$$

In fact, the cubature rule $Q_{\mathbf{X}}^m$ is the implicit integration of the least squares solution $S_{\mathbf{X}}^m f := \sum_{k=1}^{m-1} c_k \eta_k$, cf. (3.2), $\mathbf{c} := (\mathbf{L}_m^* \mathbf{L}_m)^{-1} \mathbf{L}_m^* \mathbf{f}$, since

$$Q_{\mathbf{X}}^m f = \sum_{j=1}^n q_j f(\mathbf{x}^j) = \mathbf{b}^\top (\mathbf{L}_m^* \mathbf{L}_m)^{-1} \mathbf{L}_m^* \mathbf{f} = \sum_{k=1}^{m-1} c_k \int_D \eta_k \, d\mu_D = \int_D S_{\mathbf{X}}^m f \, d\mu_D.$$

Using this, we give upper bounds on the integration error caused by $Q_{\mathbf{X}}^m$.

Theorem 7.1. *Let μ_D be a measure on D such that $L_1(D, \varrho_D) \hookrightarrow L_1(D, \mu_D)$. Denote with*

$$C_{\varrho, \mu} := \| \text{Id} : L_1(D, \varrho_D) \rightarrow L_1(D, \mu_D) \| < \infty$$

the norm of the embedding. Under the same assumptions as in Theorem 5.5 it holds for fixed $\delta > 0$

$$\mathbb{P} \left(\sup_{\|f\|_{H(K)} \leq 1} \left| \int_D f \, d\mu_D - Q_{\mathbf{X}}^m f \right|^2 \leq \frac{29 C_{\varrho, \mu}^2}{\delta} \max \left\{ \sigma_m^2, \frac{\log(8n)}{n} T(m) \right\} \right) \geq 1 - 3\delta.$$

Proof. Using the embedding relation of the different L_1 -spaces we have

$$\begin{aligned} \left| \int_D f \, d\mu_D - Q_{\mathbf{X}}^m f \right| &= \left| \int_D f - S_{\mathbf{X}}^m f \, d\mu_D \right| \leq \int_D |f - S_{\mathbf{X}}^m f| \, d\mu_D \\ &\leq C_{\varrho, \mu} \int_D |f - S_{\mathbf{X}}^m f| \, d\varrho_D \\ &\leq C_{\varrho, \mu} \|f - S_{\mathbf{X}}^m f\|_{L_2(D, \varrho_D)}. \end{aligned} \quad (7.2)$$

We conclude the proof using Corollary 5.6. □

In Theorem 7.1 we assume that the kernel is bounded, i.e., $\sup_{\mathbf{x} \in D} K(\mathbf{x}, \mathbf{x}) < \infty$. However, as we will see below it is enough to assume $\int_D K(\mathbf{x}, \mathbf{x}) \varrho_D(d\mathbf{x}) < \infty$ and that ϱ_D is a finite measure. We define the following modified (reweighted) cubature formula

$$\begin{aligned}\tilde{Q}_{\mathbf{X}}^m f &:= \int_D \tilde{S}_{\mathbf{X}}^m f(\mathbf{x}) \varrho_D(\mathbf{x}) = \tilde{\mathbf{q}}^\top \mathbf{f} \\ &= (\mathbf{D}_{\varrho_m} \overline{\tilde{\mathbf{L}}_m} \overline{(\tilde{\mathbf{L}}_m^* \tilde{\mathbf{L}}_m)^{-1}} \mathbf{b})^\top \mathbf{f} = \mathbf{b}^\top (\tilde{\mathbf{L}}_m^* \tilde{\mathbf{L}}_m)^{-1} \tilde{\mathbf{L}}_m^* \mathbf{D}_{\varrho_m} \mathbf{f},\end{aligned}$$

where \mathbf{f} is the vector of function samples $(f(\mathbf{x}^1), \dots, f(\mathbf{x}^n))$, the vector \mathbf{b} is given as in (7.1), $\tilde{\mathbf{L}}_m$ as in Algorithm 2, and

$$\mathbf{D}_{\varrho_m} = \text{diag} \left(1/\sqrt{\varrho_m(\mathbf{x}^1)}, \dots, 1/\sqrt{\varrho_m(\mathbf{x}^n)} \right).$$

Here $\varrho_m(\cdot)$ is given by (5.12) above and depends on the spectral properties of the kernel. Note that we simply put the respective entry to 0 in \mathbf{D}_{ϱ_m} if $\varrho_m(\mathbf{x}^n)$ happens to be zero.

Theorem 7.2. *If ϱ_D denotes a finite measure and $\int_D K(\mathbf{x}, \mathbf{x}) \varrho_D(d\mathbf{x}) < \infty$, then*

$$\sup_{\|f\|_{H(K)} \leq 1} \left| \int_D f(\mathbf{x}) \varrho_D(d\mathbf{x}) - \tilde{Q}_{\mathbf{X}}^m(f) \right|^2 \leq \frac{50}{\delta} \varrho_D(D) \max \left\{ \sigma_m^2, \frac{\log(8n)}{n} \sum_{j=m}^{\infty} \sigma_j^2 \right\} \quad (7.3)$$

holds with probability larger than $1 - 3\delta$ if the n nodes in \mathbf{X} are drawn i.i.d. according to (5.11) above.

Proof. Using the bound on the $L_2(D, \varrho_D)$ error from Theorem 5.8 together with (7.2) above, we obtain (7.3) with high probability if the nodes \mathbf{X} are sampled according to the measure (5.11). \square

Remark 7.3. The previous result essentially improves on the bound in [2, Thm. 1] if the singular values decay polynomially. Note that we assume that ϱ_D is a finite measure and $\int_D K(\mathbf{x}, \mathbf{x}) \varrho_D(d\mathbf{x}) < \infty$. In [2, Thm. 1] a logarithmic oversampling is not required, however the error bounds are worse by a factor n , which is substantial, e.g., in the case of the Sobolev kernel, see Section 8.2.

8 Hyperbolic cross Fourier regression

In the sequel we are interested in the recovery of functions from periodic Sobolev spaces. That is, we consider functions on the torus $\mathbb{T}^d \simeq [0, 1]^d$ where opposite points are identified. Note that the unit cube $[0, 1]^d$ is preferred here since it has Lebesgue measure 1 and is therefore a probability space. We could have also worked with $[0, 2\pi]^d$ and the Lebesgue measure (which can be made a probability measure by a d -dependent rescaling). The general error bounds for the recovery error given below (in terms of $(\sigma_j)_{j \in \mathbb{N}}$ like in Theorem 8.2) are not affected by this rescaling since the sequence $(\sigma_j)_{j \in \mathbb{N}}$ then also changes. However, some of the preasymptotic estimates for the $(\sigma_j)_{j \in \mathbb{N}}$ are sensitive with respect to a different domain as the results in Krieg [32] point out.

For $\alpha \in \mathbb{N}$ we define the space $H_{\text{mix}}^\alpha(\mathbb{T}^d)$ as the Hilbert space with the inner product

$$\langle f, g \rangle_{H_{\text{mix}}^{\alpha,*}} := \sum_{\mathbf{j} \in \{0, \alpha\}^d} \langle D^{(\mathbf{j})} f, D^{(\mathbf{j})} g \rangle_{L_2(\mathbb{T}^d)}. \quad (8.1)$$

Defining the weight

$$w_{\alpha,*}(k) = (1 + (2\pi|k|)^{2\alpha})^{1/2}, \quad k \in \mathbb{Z}, \quad (8.2)$$

and the univariate kernel function

$$K_{\alpha,*}^1(x, y) := \sum_{k \in \mathbb{Z}} \frac{\exp(2\pi i k(y - x))}{w_{\alpha,*}(k)^2}, \quad x, y \in \mathbb{T},$$

directly leads to

$$K_{\alpha,*}^d(\mathbf{x}, \mathbf{y}) := K_{\alpha,*}^1(x_1, y_1) \otimes \cdots \otimes K_{\alpha,*}^1(x_d, y_d), \quad \mathbf{x}, \mathbf{y} \in \mathbb{T}^d, \quad (8.3)$$

which is a reproducing kernel for $H_{\text{mix}}^\alpha(\mathbb{T}^d)$. The Fourier series representation of $K_{\alpha,*}^d(\mathbf{x}, \mathbf{y})$ is specified by

$$K_{\alpha,*}^d(\mathbf{x}, \mathbf{y}) := \sum_{\mathbf{k} \in \mathbb{Z}^d} \frac{\exp(2\pi i \mathbf{k} \cdot (\mathbf{y} - \mathbf{x}))}{w_{\alpha,*}(k_1)^2 \cdots w_{\alpha,*}(k_d)^2} = \sum_{\mathbf{k} \in \mathbb{Z}^d} \frac{\exp(2\pi i \mathbf{k} \cdot (\mathbf{y} - \mathbf{x}))}{\prod_{j=1}^d (1 + (2\pi|k_j|)^{2\alpha})}, \quad \mathbf{x}, \mathbf{y} \in \mathbb{T}^d.$$

In particular, for any $f \in H_{\text{mix}}^\alpha(\mathbb{T}^d)$ we have

$$f(\mathbf{x}) = \langle f, K_{\alpha,*}^d(\mathbf{x}, \cdot) \rangle_{H_{\text{mix}}^{\alpha,*}}.$$

The kernel defined in (8.3) associated to the inner product (8.1) can be extended to the case of fractional smoothness $s > 0$ replacing α by s in (8.2)–(8.3) which in turn leads to the inner product

$$\langle f, g \rangle_{H_{\text{mix}}^{s,*}} := \sum_{\mathbf{k} \in \mathbb{Z}^d} \hat{f}_{\mathbf{k}} \overline{\hat{g}_{\mathbf{k}}} \prod_{j=1}^d w_{s,*}(k_j)^2$$

and the corresponding norm

$$\|f\|_* := \|f\|_{H_{\text{mix}}^{s,*}} := \left(\sum_{\mathbf{k} \in \mathbb{Z}^d} |\hat{f}_{\mathbf{k}}|^2 \prod_{j=1}^d w_{s,*}(k_j)^2 \right)^{1/2}.$$

The (ordered) sequence $(\lambda_j^*)_{j=1}^\infty$ of eigenvalues of the corresponding mapping $W = \text{Id}^* \circ \text{Id}$, where $\text{Id}: H(K_{s,*}^d) \rightarrow L_2(\mathbb{T}^d)$ is the non-increasing rearrangement of the numbers

$$\left\{ \lambda_{\mathbf{k}}^* := \prod_{j=1}^d w_{s,*}(k_j)^2 = \prod_{j=1}^d (1 + (2\pi|k_j|)^{2s})^{-1} : \mathbf{k} \in \mathbb{Z}^d \right\}.$$

The corresponding orthonormal system $(e_j^*)_{j=1}^\infty$ in $H(K_{s,*}^d)$ is given by

$$\left\{ e_{\mathbf{k}}(\mathbf{x}) := \exp(2\pi i \mathbf{k} \cdot \mathbf{x}) \prod_{j=1}^d (1 + (2\pi|k_j|)^{2s})^{-1/2} : \mathbf{k} \in \mathbb{Z}^d \right\}.$$

Consequently, the orthonormal system $(\eta_j^*)_{j=1}^\infty$ in $L_2(\mathbb{T}^d)$ is the properly ordered classical Fourier system $\eta_{\mathbf{k}}(\mathbf{x}) = \exp(2\pi i \mathbf{k} \cdot \mathbf{x})$. This directly implies the following behavior of the quantity $N(m)$ defined in (2.7). It holds

$$N(m) = m - 1 \quad \text{and} \quad T_*(m) = \sum_{j=m}^\infty \lambda_j^* = \sum_{j=m}^\infty (\sigma_j^*)^2.$$

Remark 8.1. It is possible to define a smoothness vector $\mathbf{s} = (s_1, \dots, s_d)$ to emphasize different smoothness in different coordinate directions. Such kernels will be denoted with $K_{\mathbf{s}}(\mathbf{x}, \mathbf{y})$. In [36] the authors establish preasymptotic error bounds which can be used for the least squares analysis as we will see below.

Recent estimates in [34] allow for determining uniform recovery guarantees with preasymptotic error bounds. For this study, we need to change the kernel weight to a less natural, but for preasymptotic considerations more convenient structure of the weight

$$w_{s,\#}(k) = (1 + 2\pi|k|)^s, \quad k \in \mathbb{Z}, \quad (8.4)$$

$s > 0$. As a consequence, the univariate kernel

$$K_{s,\#}^1(x, y) := \sum_{k \in \mathbb{Z}} \frac{\exp(2\pi i k(y - x))}{w_{s,\#}(k)^2}, \quad x, y \in \mathbb{T},$$

as well as the tensor product kernel

$$K_{s,\#}^d(\mathbf{x}, \mathbf{y}) := \sum_{\mathbf{k} \in \mathbb{Z}^d} \frac{\exp(2\pi i \mathbf{k} \cdot (\mathbf{y} - \mathbf{x}))}{\prod_{j=1}^d (1 + 2\pi|k_j|)^s}, \quad \mathbf{x}, \mathbf{y} \in \mathbb{T}^d,$$

changes and has modified Fourier series expansions. Of course, the weight $w_{s,\#}$ yields an equivalent norm

$$\|f\|_{\#} := \|f\|_{H_{\text{mix}}^{s,\#}(\mathbb{T}^d)} = \left(\sum_{\mathbf{k} \in \mathbb{Z}^d} |\hat{f}_{\mathbf{k}}|^2 \prod_{j=1}^d (1 + 2\pi|k_j|)^{2s} \right)^{1/2}$$

in the space $H_{\text{mix}}^s(\mathbb{T}^d)$. However, from a complexity theoretic point of view, it is worth noting the difference of both approaches. The respective unit balls belonging to both norms differ significantly since the equivalence constants for both norms may depend badly on d . Moreover, we stress on the fact that the non-increasing rearrangements of the eigenvalues $(\lambda_j^{\#})_{j=1}^{\infty}$ of the mapping $W = \text{Id}^* \circ \text{Id}$, where $\text{Id}: H(K_{s,\#}^d) \rightarrow L_2(\mathbb{T}^d)$, differs from $(\lambda_j^*)_{j=1}^{\infty}$ since the associated mappings $j \rightarrow \mathbf{k}$ do not coincide. Accordingly, the sampling operators $S_{\mathbf{x}}^{m,*}$ and $S_{\mathbf{x}}^{m,\#}$ are defined with respect to the different non-increasing rearrangements of the basis functions $\eta_{\mathbf{k}}$, and thus, may also differ.

Despite the fact that there are many more different equivalent norms for $H_{\text{mix}}^s(\mathbb{T}^d)$, we will use only the mentioned ones in order to apply advantageous known upper bounds on the eigenvalues λ_j^* and $\lambda_j^{\#}$. In the regime of interest here, the recovery of periodic functions with mixed smoothness, we even have preasymptotic bounds for these eigenvalues/singular values available (see [35, 34, 32]). Note that theoretically everything is known about the singular values σ_m^{\square} , $\square \in \{*, \#\}$, since the behavior of this sequence is determined by the non-increasing rearrangement of the reciprocals of the tensor product weights $\prod_{j=1}^d w_{s,\square}(k_j)$, cf. (8.2) and (8.4). The analysis of the rearrangement of multi-indexed sequences has been revealed that the upper bound

$$\sigma_m^{\#} \leq \min \left(1, \frac{16}{3m} \right)^{\frac{s}{1+\log_2 d}} \leq \left(\frac{16}{3m} \right)^{\frac{s}{1+\log_2 d}} \quad (8.5)$$

holds for all $m \in \mathbb{N}$, cf. [34, Thm. 4.1].

One may argue that the kernel $K_{s,\#}^d$ is less “natural” than the kernel $K_{s,*}^d$. For this purpose we use the observation by Krieg [32], which yields the preasymptotic bound

$$\sigma_m^* \leq \left(\frac{1.26}{m} \right)^{\frac{1.83 \cdot s}{4 + \log_2(d)}}$$

in the range $2 \leq m \leq 3^d$, where the exponent scales also linearly in s .

8.1 Uniform recovery of functions from $H_{\text{mix}}^s(\mathbb{T}^d)$

First, we consider the asymptotic behavior of the sampling error caused by the presented least squares approach. Since the asymptotic bounds on λ_j^* and $\lambda_j^\#$ differ only by constants, that we do not specify explicitly, we study both cases collectively. For $\square \in \{*, \#\}$, the asymptotic behavior of the sequence $(\sigma_j^\square)_{j=1}^\infty$ for the embedding $\text{Id} : H(K_{s,\square}^d) \rightarrow L_2(\mathbb{T}^d)$ has been known for a long time, cf. [17, Chapt. 4] and the references therein. There is a constant \tilde{C}_d^\square which depends exponentially on d such that

$$\sigma_m^\square \leq \tilde{C}_d^\square m^{-s} (\log m)^{s(d-1)}, \quad m \in \mathbb{N}. \quad (8.6)$$

As a direct consequence of Corollary 5.6 we determine explicit error bounds as well as the asymptotic error behavior (8.7), where the latter holds for all equivalent norms in $H_{\text{mix}}^s(\mathbb{T}^d)$. Please note that the constants C_d^\square depend heavily on the specific norm. Moreover, a statement similar to (8.7) can be also obtained with the technique in [33].

Theorem 8.2. *Let $d \in \mathbb{N}, s > 1/2, \delta > 0$ and $n \in \mathbb{N}$ be given. Choose $m \in \mathbb{N}, m \geq 2$,*

$$m \leq \left\lfloor \frac{n}{48(\sqrt{2} \log(2n) - \log \delta)} \right\rfloor + 1$$

and draw the n sampling nodes in \mathbf{X} uniformly i.i.d. at random in \mathbb{T}^d . Then, we achieve

$$\mathbb{P} \left(\sup_{\|f\|_\square \leq 1} \|f - S_{\mathbf{X}}^{m,\square} f\|_{L_2(\mathbb{T}^d)}^2 \leq \frac{29}{\delta} \max \left\{ (\sigma_m^\square)^2, \frac{\log(8n)}{n} \sum_{k=m}^\infty (\sigma_k^\square)^2 \right\} \right) \geq 1 - 3\delta.$$

In particular, there is a constant $C_d^\square > 0$ depending on d such that for $0 < \delta < 1/3$ it holds

$$\mathbb{P} \left(\sup_{\|f\|_\square \leq 1} \|f - S_{\mathbf{X}}^{m,\square} f\|_{L_2(\mathbb{T}^d)} \leq \frac{C_d^\square}{\sqrt{\delta}} n^{-s} (\log n)^{sd} \right) \geq 1 - 3\delta. \quad (8.7)$$

Proof. The result follows from Corollary 5.6 and our specific situation where $B = 1$, $N(m) = m - 1$ and $T_\square(m) = \sum_{k=m}^\infty (\sigma_k^\square)^2$. We estimate $T_\square(m)$ using [12]

$$T_\square(m) := \sup_{\mathbf{x} \in D} \sum_{k=m}^\infty |e_k(\mathbf{x})|^2 = \sum_{k=m}^\infty (\sigma_k^\square)^2 \asymp m^{-2(s-1/2)} (\log m)^{2(d-1)s} \asymp m (\sigma_m^\square)^2.$$

Hence, the right-hand side in (5.9) can be bounded from above by a constant times σ_m^\square , which behaves as $n^{-s} (\log n)^{sd}$. \square

In addition, we investigate the preasymptotic error behavior using the aforementioned estimates (8.5) on the singular values $\sigma_m^\#$ that belongs to $\text{Id}: H(K_{s,\#}^d) \rightarrow L_2(\mathbb{T}^d)$. Since the upper bounds have been proven only for this specific type of mappings, the following results, in particular the explicitly determined constants, may only hold for RKHS with weight functions $w_s(\mathbf{k}) \geq \prod_{j=1}^d (1 + |k_j|)^s$, which is fulfilled for $w_{s,\#}$.

Theorem 8.3. *Let $d \in \mathbb{N}$, $s > (1 + \log_2 d)/2$, $0 < \delta < 1$, and $n \in \mathbb{N}$ such that*

$$m := \left\lfloor \frac{n}{48(\sqrt{2} \log(2n) - \log \delta)} \right\rfloor + 1 \quad (8.8)$$

is at least 2. Drawing the n sampling nodes in \mathbf{X} uniformly i.i.d. at random in \mathbb{T}^d yields

$$\mathbb{E} \left(\sup_{\|f\|_\# \leq 1} \|f - S_{\mathbf{X}}^{m,\#} f\|_{L_2(\mathbb{T}^d)}^2 \middle| \|\mathbf{H}_{m,\#} - \mathbf{I}_m\| \leq 1/2 \right) \leq \frac{29}{6(1-\delta)} \frac{2s}{2s-1-\log_2 d} \left(\frac{16}{3m} \right)^{\frac{2s}{1+\log_2 d}}.$$

Proof. We apply Theorem 5.5 and take into account that we choose

$$m := \left\lfloor \frac{n}{48(\sqrt{2} \log(2n) - \log \delta)} \right\rfloor + 1 \leq \frac{n}{48\sqrt{2} \log(2n)} + 1$$

for large enough n . Furthermore, we set $c := 16/3$, $\beta := 2s/(1 + \log_2 d)$ and estimate $T_\#(m)$, see (8.5),

$$\begin{aligned} T_\#(m) &:= \sup_{\mathbf{x} \in D} \sum_{k=m}^{\infty} |e_k(\mathbf{x})|^2 = \sum_{k=m}^{\infty} (\sigma_k^\#)^2 \leq c^\beta \sum_{k=m}^{\infty} k^{-\beta} \\ &\leq c^\beta \left(m^{-\beta} + \frac{1}{\beta-1} m^{-\beta+1} \right) \leq c^\beta \frac{\beta}{\beta-1} m^{-\beta+1}. \end{aligned}$$

Taking (5.4) into account, we bound

$$\begin{aligned} &\mathbb{E} \left(\sup_{\|f\|_\# \leq 1} \|f - S_{\mathbf{X}}^{m,\#} f\|_{L_2(\mathbb{T}^d)}^2 \middle| \|\mathbf{H}_{m,\#} - \mathbf{I}_m\| \leq 1/2 \right) \\ &\leq \frac{c^\beta m^{-\beta}}{1-\delta} \left(3 + 8 C_R^2 \frac{m \log(8n)}{n} \frac{\beta}{\beta-1} + 4 C_R \sqrt{\frac{m \log(8n)}{n} \frac{\beta}{\beta-1}} \right) \\ &\leq \frac{\beta c^\beta m^{-\beta}}{(1-\delta)(\beta-1)} \left(3 + 8 C_R^2 b + 4 C_R \sqrt{b} \right), \end{aligned}$$

where $b := \frac{\log(8n)}{48\sqrt{2} \log(2n)} + \frac{\log(8n)}{n}$. The term in the brackets is monotonically decreasing in n . We stress that the last estimates are reasonable for $m \geq 2$ and thus, we need at least $n \geq 464$. This choice of n leads to an upper bound of the term within the brackets which is $29/6$. Thus, for $m \geq 2$, the estimate

$$\mathbb{E} \left(\sup_{\|f\|_\# \leq 1} \|f - S_{\mathbf{X}}^{m,\#} f\|_{L_2(\mathbb{T}^d)}^2 \middle| \|\mathbf{H}_{m,\#} - \mathbf{I}_m\| \leq 1/2 \right) \leq \frac{29 \beta c^\beta m^{-\beta}}{6(1-\delta)(\beta-1)}$$

holds and the assertion follows. \square

Similar to Corollary 5.6, we apply Markov's inequality to get a lower bound on the success probability of the randomly chosen sampling set.

Corollary 8.4. *Under the same assumptions as in Theorem 8.3 it holds*

$$\begin{aligned} \mathbb{P}\left(\sup_{\|f\|_{\#} \leq 1} \|f - S_{\mathbf{X}}^{m,\#} f\|_{L_2(\mathbb{T}^d)}^2\right. \\ \leq \frac{29}{6\delta} \frac{2s}{2s-1-\log_2 d} \left(\frac{256(\sqrt{2}\log(2n) - \log\delta)}{n}\right)^{\frac{2s}{1+\log_2 d}} \\ \geq \mathbb{P}\left(\sup_{\|f\|_{\#} \leq 1} \|f - S_{\mathbf{X}}^{m,\#} f\|_{L_2(\mathbb{T}^d)}^2 \leq \frac{29}{6\delta} \frac{2s}{2s-1-\log_2 d} \left(\frac{16}{3m}\right)^{\frac{2s}{1+\log_2 d}}\right) \geq 1 - 3\delta. \end{aligned}$$

Proof. We follow the argumentation in the proof of Corollary 5.6 using the inequality

$$m^{-1} \leq \frac{48(\sqrt{2}\log(2n) - \log\delta)}{n}.$$

□

Example 8.5. For $s = 5$ and $d = 16$, we would like to fulfill

$$\mathbb{P}\left(\sup_{\|f\|_{\#} \leq 1} \|f - S_{\mathbf{X}}^{m,\#} f\|_{L_2(\mathbb{T}^d)} \leq 0.1\right) \geq 0.99,$$

i.e., with $\delta = 1/300$ we choose $m := 2873$ smallest possible such that

$$2900 \left(\frac{16}{3m}\right)^2 \leq 0.01$$

holds. Clearly, for n such that $m-1 = 2872 \leq \frac{n}{48(\sqrt{2}\log(2n) - \log\delta)}$ holds, we observe the desired estimate. We choose the smallest possible $n := 3879\,166$.

8.2 Numerical integration of periodic functions

In [50, Sec. 4.2], the author discussed the construction of stable cubature weights for the approximation of the integral

$$I(f) := \int_{\mathbb{T}^d} f(\mathbf{x}) d\mathbf{x} \approx Q_{\mathbf{X}}^m f := \sum_{j=1}^n q_j f(\mathbf{x}^j)$$

for functions from periodic Sobolev spaces with dominating mixed smoothness. The integration nodes \mathbf{X} are drawn uniformly and independently at random from \mathbb{T}^d . Below, in Corollary 8.6, the cubature rule $Q_{\mathbf{X}}^m$ is fixed for the whole class. The result in [50, Thm. 4.5] bounds the worst-case integration error from above by

$$\lesssim_{d,s} n^{-s+1/2} (\log n)^{sd-1/2}.$$

Corresponding numerical tests promise better behavior of the integration error, cf. [50, Rem. 4.6]. Our theoretical results of this section confirm that the optimal main rate of the presented approach in [50] is n^{-s} , in particular we obtain the upper bound $\lesssim_{d,s} n^{-s} (\log n)^{sd}$ for this specific setting. We achieve the following statement on the worst-case integration error.

Corollary 8.6. *Let $d \in \mathbb{N}, s > 1/2$ and $0 < \delta < 1/3$. We choose $n \in \mathbb{N}$ such that m as stated in (8.8) is at least 2. Drawing the n sampling nodes in \mathbf{X} uniformly i.i.d. at random from \mathbb{T}^d , we put the cubature weight vector \mathbf{q} to be the first column of $\overline{\mathbf{L}_m(\mathbf{L}_m^* \mathbf{L}_m)^{-1}}$. Then, with probability at least $1 - 3\delta$, we obtain*

$$\sup_{\|f\|_{\square} \leq 1} |I(f) - Q_{\mathbf{X}}^m f| \leq \sqrt{\frac{29}{\delta}} \max \left\{ \sigma_m^{\square}, \sqrt{\frac{\log(8n)}{n} \sum_{k=m}^{\infty} (\sigma_k^{\square})^2} \right\}.$$

In particular, there is a constant $C_d^{\square} > 0$ depending on d such that for $0 < \delta < 1/3$ it holds with probability $1 - 3\delta$

$$\sup_{\|f\|_{\square} \leq 1} |I(f) - Q_{\mathbf{X}}^m f| \leq \frac{C_d^{\square}}{\sqrt{\delta}} n^{-s} (\log n)^{sd}. \quad (8.9)$$

Proof. We apply Theorem 7.1 with $\mu_{\mathbb{T}^d} = \varrho_{\mathbb{T}^d} \equiv 1$ followed by (8.6). \square

Remark 8.7. By the same reasoning, the result in Theorem 8.3 transfers almost literally to the integration problem. In fact, having $s > (1 + \log_2 d)/2$ we see a non-trivial preasymptotic behavior. The above bounds show that this method based on random points competes with most of the quasi-Monte-Carlo methods studied in the literature, see [21, pp. 195, 247].

9 Hyperbolic wavelet regression

The following scenario of replacing the Fourier system by dyadic wavelets has already been investigated by Bohn [5, Sec. 5.5.2], [6] using piecewise linear prewavelets. Here we use orthogonal wavelets and will improve the result in [5] in two directions. First, we remove a d -dependent log-factor and second, our result holds for the whole class and not just one individual function, i.e., we control the worst-case error. It is worth mentioning, that we only loose a log-factor which is independent of d compared to the benchmark result in [20].

Let us start with the necessary definitions since we are now in a non-periodic setting. For $s > 0$ let us define the space $H_{\text{mix}}^s(\mathbb{R}^d)$ as the collection of all functions from $L_2(\mathbb{R}^d)$ such that

$$\|f\|_{H_{\text{mix}}^s(\mathbb{R}^d)} = \left\| \left(\prod_{i=1}^d (1 + |y_i|)^s \right) \mathcal{F}f(\mathbf{y}) \right\|_{L_2(\mathbb{R}^d)} < \infty.$$

Here \mathcal{F} denotes the Fourier transform on \mathbb{R}^d given by

$$\mathcal{F}f(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^d} \int_{\mathbb{R}^d} f(\mathbf{y}) \exp(-i \mathbf{y} \cdot \mathbf{x}) d\mathbf{y}, \quad \mathbf{x} \in \mathbb{R}^d.$$

It is well-known, that $H_{\text{mix}}^s(\mathbb{R}^d)$ can be characterized using hyperbolic wavelets. Let $(\psi_{j,k})_{j \in \mathbb{N}_0, k \in \mathbb{Z}}$ be a univariate orthonormal wavelet system (if $j = 0$ then $\psi_{0,k}$ denotes the orthogonal scaling function). Then we denote with

$$\psi_{\mathbf{j}, \mathbf{k}}(\mathbf{x}) := \psi_{i_1, k_1}(x_1) \cdots \psi_{i_d, k_d}(x_d), \quad \mathbf{x} \in \mathbb{R}^d, \quad \mathbf{j} \in \mathbb{N}_0^d, \quad \mathbf{k} \in \mathbb{Z}^d,$$

the corresponding hyperbolic wavelet basis in $L_2(\mathbb{R}^d)$. For our analysis we need that the univariate wavelet is a compactly supported wavelet, which means that $\psi_{j,k}$ is supported

“near” the interval $[k2^{-j}, (k+1)2^{-j}]$. If the wavelet basis has sufficient smoothness and vanishing moments, then $f \in H_{\text{mix}}^s(\mathbb{R}^d)$ holds if and only if

$$\left(\sum_{\mathbf{j} \in \mathbb{N}_0^d} \sum_{\mathbf{k} \in \mathbb{Z}^d} 2^{2\|\mathbf{j}\|_1 s} |\langle f, \psi_{\mathbf{j}, \mathbf{k}} \rangle|^2 \right)^{1/2} < \infty.$$

This leads to the norm equivalence

$$\|f\|_{H_{\text{mix}}^s(\mathbb{R}^d)} \asymp \left(\sum_{\mathbf{j} \in \mathbb{N}_0^d} \sum_{\mathbf{k} \in \mathbb{Z}^d} 2^{2\|\mathbf{j}\|_1 s} |\langle f, \psi_{\mathbf{j}, \mathbf{k}} \rangle|^2 \right)^{1/2}. \quad (9.1)$$

Clearly, if $\|f\|_{H_{\text{mix}}^s(\mathbb{R}^d)} \leq 1$, then the sequence $(2^{\|\mathbf{j}\|_1 s} \langle f, \psi_{\mathbf{j}, \mathbf{k}} \rangle)_{\mathbf{j}, \mathbf{k}}$ has an ℓ_2 -norm bounded by a constant, which will be important for our later analysis.

Let us consider the unit cube $[0, 1]^d$. Let further $D_{\mathbf{j}}$ be the set of all $\mathbf{k} \in \mathbb{Z}^d$ such that the wavelet $\text{supp } \psi_{\mathbf{j}, \mathbf{k}}$ has a non-empty intersection with $[0, 1]^d$. This directly leads to the extended domain Ω given by

$$\Omega := \bigcup_{\mathbf{j} \in \mathbb{N}_0^d} \bigcup_{\mathbf{k} \in D_{\mathbf{j}}} \text{supp } \psi_{\mathbf{j}, \mathbf{k}}.$$

It holds $[0, 1]^d \subset \Omega$, and the system $(\psi_{\mathbf{j}, \mathbf{k}})_{\mathbf{j} \in \mathbb{N}_0^d, \mathbf{k} \in D_{\mathbf{j}}}$ is an orthonormal system in $L_2(\Omega)$, however not a basis. Note that Ω is still a bounded tensor domain with a measure proportional to 1 depending on the support length of the wavelet basis. It is also clear that this orthonormal system is not uniformly bounded in L_∞ .

In the sequel we want to recover functions $f \in H_{\text{mix}}^s(\mathbb{R}^d)$ on the domain $[0, 1]^d$ from samples on the slightly larger extended domain Ω in a uniform way. In other words, the discrete locations of the sampling nodes $\mathbf{X} = \{\mathbf{x}^1, \dots, \mathbf{x}^n\}$ are chosen in advance for the whole class of functions. Let us consider the operator

$$\tilde{P}_\ell f := \sum_{\|\mathbf{j}\|_1 \leq \ell} \sum_{\mathbf{k} \in D_{\mathbf{j}}} \langle f, \psi_{\mathbf{j}, \mathbf{k}} \rangle \psi_{\mathbf{j}, \mathbf{k}}, \quad \ell \in \mathbb{N},$$

which is known from *hyperbolic wavelet approximation*, see [20, 58]. The following worst-case error bound is well-known and follows directly from (9.1):

$$\sup_{\|f\|_{H_{\text{mix}}^s(\mathbb{R}^d)} \leq 1} \|f - \tilde{P}_\ell f\|_{L_2([0, 1]^d)} \asymp 2^{-s\ell}.$$

We now consider a special case of the matrix \mathbf{L}_m from (3.1), namely

$$\tilde{\mathbf{L}}_\ell := \begin{pmatrix} (\tilde{\psi}_{\mathbf{j}, \mathbf{k}}(\mathbf{x}^1))_{\|\mathbf{j}\|_1 \leq \ell, \mathbf{k} \in D_{\mathbf{j}}}^\top \\ \vdots \\ (\tilde{\psi}_{\mathbf{j}, \mathbf{k}}(\mathbf{x}^n))_{\|\mathbf{j}\|_1 \leq \ell, \mathbf{k} \in D_{\mathbf{j}}}^\top \end{pmatrix}. \quad (9.2)$$

Here, $m = m(\ell) \asymp 2^\ell \ell^{d-1}$ and the functions $\tilde{\psi}_{\mathbf{j}, \mathbf{k}} = \sqrt{|\Omega|} \psi_{\mathbf{j}, \mathbf{k}}$ enumerate the properly re-normalized wavelets $\psi_{\mathbf{j}, \mathbf{k}}, \|\mathbf{j}\|_1 \leq \ell, \mathbf{k} \in D_{\mathbf{j}}$, which is now an orthonormal system in the space $L_2(\Omega, \varrho_\Omega)$ with the probability measure $\varrho_\Omega = \frac{d\mathbf{x}}{|\Omega|}$. The n sampling nodes in \mathbf{X} i.i.d. at random according to ϱ_Ω . Note that, due to the construction, we have that $|\Omega|$ is bounded by

Algorithm 3 Hyperbolic wavelet regression.

Input: $\ell \in \mathbb{N}$,
 $n: \tilde{N}(\ell) \asymp 2^\ell \ell^{d-1} \lesssim \frac{n}{\log(n)}$,
 $\mathbf{X} = (\mathbf{x}^1, \dots, \mathbf{x}^n) \in D^n$ set of distinct sampling nodes,
 $\mathbf{f} = (f(\mathbf{x}^1), \dots, f(\mathbf{x}^n))^\top$ samples of f evaluated at the nodes from \mathbf{X} ,
such that the matrix $\tilde{\mathbf{L}}_\ell := \tilde{\mathbf{L}}_\ell(\mathbf{X})$ from (9.2)
has full (column) rank.

Solve the over-determined linear system

$$\tilde{\mathbf{L}}_\ell(c_{\mathbf{j}, \mathbf{k}})_{\mathbf{j}, \mathbf{k}} = \mathbf{f}$$

via least squares (e.g. directly or via the LSQR algorithm [52]), i.e., compute

$$(c_{\mathbf{j}, \mathbf{k}})_{\mathbf{j}, \mathbf{k}} := (\tilde{\mathbf{L}}_\ell^* \tilde{\mathbf{L}}_\ell)^{-1} \tilde{\mathbf{L}}_\ell^* \mathbf{f}.$$

Output: $\mathbf{c} = (c_{\mathbf{j}, \mathbf{k}})_{\mathbf{j}, \mathbf{k}} \in \mathbb{C}^{m(\ell)}$ coefficients of the approximant $S_{\mathbf{X}}^\ell f := \sum_{\|\mathbf{j}\|_1 \leq \ell} \sum_{\mathbf{k} \in D_{\mathbf{j}}} c_{\mathbf{j}, \mathbf{k}} \tilde{\psi}_{\mathbf{j}, \mathbf{k}}$.

a constant which depends on the chosen wavelet system. This, on the other hand, depends on the assumed mixed regularity properties of the function f , i.e., the mixed smoothness $s > 0$. The larger s is chosen, the larger the support of a properly chosen orthonormal wavelet system has to be. We propose Algorithm 3 for computing the wavelet coefficients of an approximation $S_{\mathbf{X}}^\ell f$ to f . Note, that there is a little abuse of notation since we use the wavelet level ℓ as upper index (in contrast to m , the dimension of the subspace used earlier).

Theorem 9.1. *Let $0 < \delta < 1$. Let further $s > 1/2$ and $(\psi_{\mathbf{j}, \mathbf{k}})_{\mathbf{j}, \mathbf{k}}$ be a hyperbolic and compactly supported orthonormal wavelet system such that (9.1) holds true. Then the algorithm $S_{\mathbf{X}}^\ell$ indicated in Algorithm 3 recovers any $f \in H_{\text{mix}}^s(\mathbb{R}^d)$ on $L_2([0, 1]^d)$ with high probability from $n = n(\ell)$ random samples which are drawn in advance for the whole class. Precisely,*

$$\sup_{\|f\|_{H_{\text{mix}}^s(\mathbb{R}^d)} \leq 1} \|f - S_{\mathbf{X}}^\ell f\|_{L_2([0, 1]^d)} \lesssim C_{\delta, d} 2^{-\ell s} \quad (9.3)$$

with probability larger than $1 - \delta$. The operator $S_{\mathbf{X}}^\ell$ uses $n(\ell) \asymp 2^\ell \ell^d$ many samples such that the bound in (9.3) reads as $\tilde{C}_{\delta, d} n^{-s} \log(n)^{ds}$ in terms of the number of samples.

Remark 9.2. (i) Note that the optimal operator \tilde{P}_ℓ uses $n(\ell) \asymp 2^\ell \ell^{d-1}$ wavelet coefficients. The gap between sampling recovery (Λ^{std}) and general linear approximation (Λ^{all}), see e.g. [17], [46, 47, 49], is reduced to a log-factor, which is independent of d .

(ii) The matrix defined in (9.2) is rather sparse. It has $n \asymp 2^\ell \ell^d$ rows and $m \asymp 2^\ell \ell^{d-1}$ columns. In every row we have only $\#\{\|\mathbf{j}\|_1 \leq \ell\} \asymp \ell^d$ many non-zero entries. This gives an additional acceleration for the least squares algorithm since matrix vector multiplications are cheap in this situation.

Proof of Theorem 9.1. We follow the proof of Theorem 5.5. Let \mathbf{X} be a set of n randomly drawn nodes from Ω according to ϱ_Ω . If the number n of samples satisfies (5.1), that is

$$\tilde{N}(\ell) := \sup_{\mathbf{x} \in \Omega} \sum_{\|\mathbf{j}\|_1 \leq \ell} \sum_{\mathbf{k} \in D_{\mathbf{j}}} \left| \tilde{\psi}_{\mathbf{j}, \mathbf{k}}(\mathbf{x}) \right|^2 \lesssim \frac{n}{\log n - \log \delta}, \quad (9.4)$$

then

$$\left\| (\tilde{\mathbf{L}}_\ell^* \tilde{\mathbf{L}}_\ell)^{-1} \tilde{\mathbf{L}}_\ell^* \right\| \leq \sqrt{\frac{2}{n}}$$

is satisfied with probability larger than $1 - \delta$. Let \mathbf{X} be such that this is the case. Then we estimate

$$\begin{aligned} \|f - S_{\mathbf{X}}^\ell f\|_{L_2([0,1]^d)} &\leq \|f - \tilde{P}_\ell f\|_{L_2([0,1]^d)} + \|\tilde{P}_\ell f - S_{\mathbf{X}}^\ell f\|_{L_2([0,1]^d)} \\ &\lesssim \left\| f - \sum_{\|\mathbf{j}\|_1 \leq \ell} \sum_{\mathbf{k} \in \mathbb{Z}^d} \langle f, \psi_{\mathbf{j}, \mathbf{k}} \rangle \psi_{\mathbf{j}, \mathbf{k}} \right\|_{L_2(\mathbb{R}^d)} + \|S_{\mathbf{X}}^\ell (\tilde{P}_\ell f - f)\|_{L_2(\Omega, \varrho_\Omega)} \\ &\lesssim 2^{-\ell s} + \sqrt{\frac{2}{n}} \cdot \left[\sum_{u=1}^n \left(\sum_{\|\mathbf{j}\|_1 > \ell} \sum_{\mathbf{k} \in \mathbb{Z}^d} \overline{\langle f, \psi_{\mathbf{j}, \mathbf{k}} \rangle} \psi_{\mathbf{j}, \mathbf{k}}(\mathbf{x}^u) \right) \left(\sum_{\|\mathbf{j}\|_1 > \ell} \sum_{\mathbf{k} \in \mathbb{Z}^d} \langle f, \psi_{\mathbf{j}, \mathbf{k}} \rangle \psi_{\mathbf{j}, \mathbf{k}}(\mathbf{x}^u) \right) \right]^{1/2} \\ &\lesssim 2^{-\ell s} + \sqrt{\frac{2}{n}} \cdot \left[\sum_{\|\mathbf{j}'\|_1 > \ell} \sum_{\mathbf{k}' \in \mathbb{Z}^d} \sum_{\|\mathbf{j}\|_1 > \ell} \sum_{\mathbf{k} \in \mathbb{Z}^d} \right. \\ &\quad \left. 2^{\|\mathbf{j}'\|_1 s} 2^{\|\mathbf{j}\|_1 s} \overline{\langle f, \psi_{\mathbf{j}', \mathbf{k}'} \rangle} \langle f, \psi_{\mathbf{j}, \mathbf{k}} \rangle \sum_{u=1}^n \frac{\overline{\psi_{\mathbf{j}', \mathbf{k}'}(\mathbf{x}^u)}}{2^{\|\mathbf{j}'\|_1 s}} \frac{\psi_{\mathbf{j}, \mathbf{k}}(\mathbf{x}^u)}{2^{\|\mathbf{j}\|_1 s}} \right]^{1/2} \\ &\lesssim 2^{-\ell s} + \sqrt{\frac{2}{n}} \| (2^{\|\mathbf{j}\|_1 s} \langle f, \psi_{\mathbf{j}, \mathbf{k}} \rangle)_{\mathbf{j}, \mathbf{k}} \|_2 \|\tilde{\Phi}_\ell\| \\ &\lesssim 2^{-\ell s} + \sqrt{\frac{2}{n}} \|\tilde{\Phi}_\ell\|, \end{aligned}$$

where $\tilde{\Phi}_\ell$ is defined similar as in Proposition 5.4. This time we put

$$\tilde{\Phi}_\ell := \begin{pmatrix} (2^{-\|\mathbf{j}\|_1 s} \psi_{\mathbf{j}, \mathbf{k}}(\mathbf{x}^1))_{\|\mathbf{j}\|_1 > \ell, \mathbf{k} \in D_{\mathbf{j}}}^\top \\ \vdots \\ (2^{-\|\mathbf{j}\|_1 s} \psi_{\mathbf{j}, \mathbf{k}}(\mathbf{x}^n))_{\|\mathbf{j}\|_1 > \ell, \mathbf{k} \in D_{\mathbf{j}}}^\top \end{pmatrix}.$$

Let us define the quantity

$$\tilde{T}(\ell) := \sup_{\mathbf{x} \in \Omega} \sum_{\|\mathbf{j}\|_1 > \ell} \sum_{\mathbf{k} \in \mathbb{Z}^d} 2^{-2\|\mathbf{j}\|_1 s} |\psi_{\mathbf{j}, \mathbf{k}}(\mathbf{x})|^2,$$

which goes along the lines of Proposition 5.4. Then we get with literally the same arguments

$$\mathbb{E} \|\tilde{\Phi}_\ell\|^2 \lesssim n \left(2^{-2\ell s} + \frac{\log n}{n} \tilde{T}(\ell) + 2^{-\ell s} \sqrt{\frac{\log n}{n} \tilde{T}(\ell)} \right). \quad (9.5)$$

Let us compute $\tilde{T}(\ell)$. Due to the compact support of the wavelet system, there are for fixed \mathbf{j} only $O(1)$ many wavelets $\psi_{\mathbf{j}, \mathbf{k}}$ such that $\psi_{\mathbf{j}, \mathbf{k}}(\mathbf{x})$ is non-zero. For those $O(1)$ wavelets, we have

$$|\psi_{\mathbf{j}, \mathbf{k}}(\mathbf{x})|^2 \lesssim 2^{\|\mathbf{j}\|_1}.$$

Hence, we get

$$\tilde{T}(\ell) \lesssim \sum_{\|\mathbf{j}\|_1 > \ell} 2^{\|\mathbf{j}\|_1(1-2s)} \lesssim 2^{\ell(1-2s)} \ell^{d-1}. \quad (9.6)$$

By the same reasoning we may estimate $\tilde{N}(\ell)$ in (9.4). Clearly n may be chosen such that

$$\tilde{N}(\ell) \asymp 2^\ell \ell^{d-1} \lesssim \frac{n}{\log(n)}. \quad (9.7)$$

Plugging (9.6) and (9.7) into (9.5) we obtain

$$\mathbb{E}\|\tilde{\Phi}_\ell\|^2 \lesssim n 2^{-2\ell s}.$$

The same standard arguments as used in Theorem 5.5 and Corollary 5.6 lead to the bound in (9.3). It remains to estimate the number of samples n depending on ℓ , see (9.7). This clearly gives $\log(n) \gtrsim \ell$ and hence $n \gtrsim 2^\ell \ell^d$ which concludes the proof. \square

10 Numerical experiments

10.1 Recovery of functions from spaces with mixed smoothness

In this section, we perform numerical tests for the hyperbolic cross Fourier regression based on random sampling nodes from Section 8, i.e., we apply Algorithm 1 to periodic test functions f from the spaces $H_{\text{mix}}^s(\mathbb{T}^d)$. In Figure 1, we visualize realizations for such random nodes in the two- and three-dimensional case.

Besides random point sets, different types of deterministic lattices have also been used for numerical integration and function recovery, see for instance [30, 31] and [9]. This motivates us to consider Frolov lattices [29] and Fibonacci lattices (cf. e.g. [60, Sec. IV.2]) in the context of this paper, see Figure 2 for examples of such lattices.

In the following, we use the weight function

$$w(\mathbf{k}) := \prod_{i=1}^d (1 + |k_i|^2)^{1/2}.$$

Note that for computational reasons we avoid the 2π in this weight w . By the reasoning after (8.5), the weights without 2π lead to a slightly slower decay of the respective singular numbers.

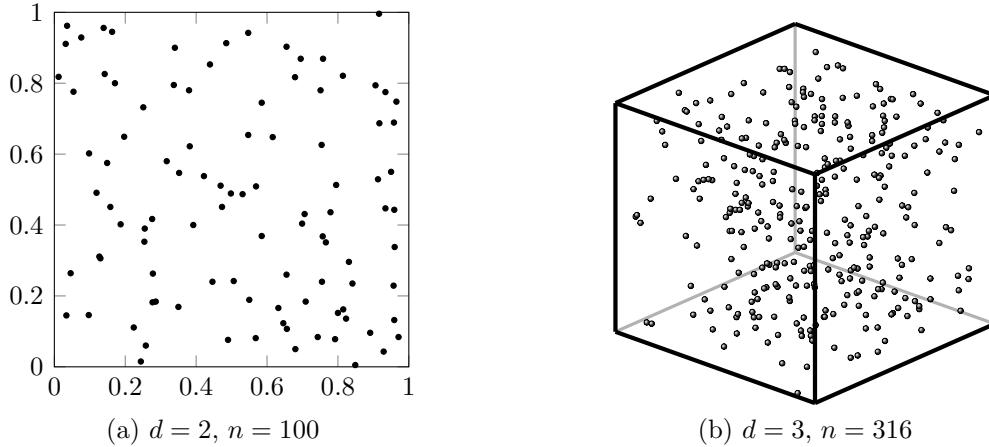


Figure 1: Realizations of random nodes for hyperbolic Fourier regression.

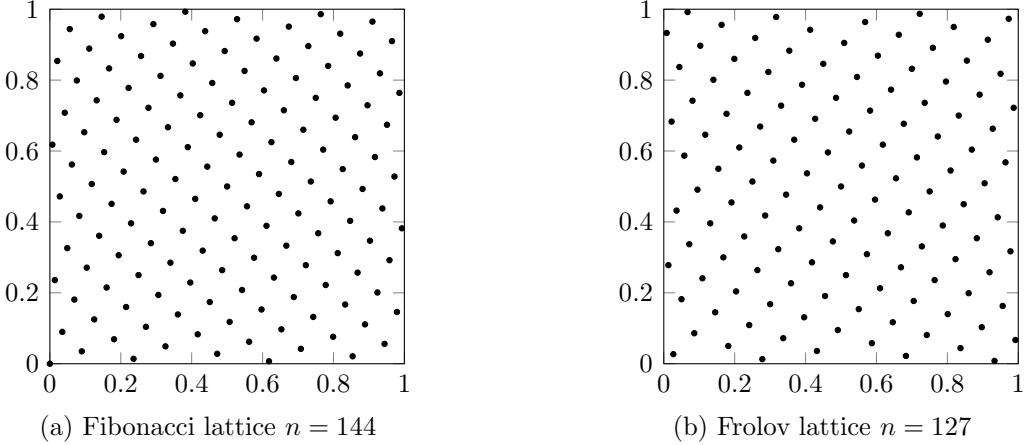


Figure 2: Examples of Fibonacci and Frolov lattices in $d = 2$ spatial dimensions.

For a given number n of samples, we use the $m = \lfloor n/(4 \log n) \rfloor$ frequencies $\mathbf{k} \in \mathbb{Z}^d$ where $w(\mathbf{k})$ is smallest, i.e., we define $I_m := \{\mathbf{k}_1, \dots, \mathbf{k}_m\} \subset \mathbb{Z}^d$, $|I_m| = m$, assuming an arrangement of \mathbf{k}_j fulfilling $w(\mathbf{k}_1) \leq w(\mathbf{k}_2) \leq \dots$. Here ties are broken in numerical order starting with the first component k_1 of \mathbf{k} until the last one k_d . Corresponding to our theoretical results, the goal is to compute a least squares approximation $S_{\mathbf{X}}^m f$ of the projection $P_{m-1} f$ of the function f to the span $\{\exp(2\pi i \mathbf{k} \cdot \mathbf{x}) : \mathbf{k} \in I_m\}$ using the n sampling values at the nodes in \mathbf{X} .

Comments on the arithmetic cost of Algorithm 1. Building the index set I_{m-1} , i.e., enumerating the basis functions $\eta_1, \dots, \eta_{m-1}$, requires

$$\leq 4 C_1 d m^2 \log m \leq C_1 d n m$$

arithmetic operations, and setting up the matrix \mathbf{L}_m requires

$$\leq C_2 d n m$$

arithmetic operations, where $C_1, C_2 > 0$ are absolute constants. Afterwards, running Algorithm 1 requires

$$\leq C_3 R n m \leq C_3 R \frac{n^2}{4 \log n}$$

arithmetic operations, where $C_3 > 0$ is an absolute constant and $R \in \mathbb{N}$ is the number of LSQR iterations. If one chooses m as in Theorem 5.5, cf. (5.1) and (5.2), the condition number of the matrix \mathbf{L}_m is $\leq \sqrt{3}$ with high probability and one obtains $R \leq 17$ for a LSQR accuracy of $\approx 10^{-8}$. Please note that we choose $m = \lfloor n/(4 \log n) \rfloor$ in our experiments, which is slightly larger than the theory (5.2) requires.

Remark 10.1. Let us compare the hyperbolic cross Fourier regression from Section 8, which uses random samples, with the single rank-1 lattice sampling approach from [30, 9], which uses highly structured deterministic sampling nodes. Up to logarithmic factors, both approaches have comparable error estimates w.r.t. the number m of basis functions and comparable arithmetic complexities. Single rank-1 lattice sampling has slightly worse recovery error estimates w.r.t. m than Algorithm 1, cf. Theorem 8.2. On the other hand, the arithmetic complexity for single rank-1 lattice sampling is slightly better. Moreover, the error estimates when using rank-1 lattices are guaranteed upper bounds, whereas the worst case

upper bounds in Section 8 hold with high probability. However, for fixed m , the used number of samples for single rank-1 lattice sampling is distinctly higher, i.e., almost quadratic compared to the approach from this paper. This results in error estimates w.r.t. the number n of used sampling values that are distinctly worse for single rank-1 lattices.

Subsequently, we consider three different test functions $f: \mathbb{T}^d \rightarrow \mathbb{R}$, where the Fourier coefficients $\hat{f}_{\mathbf{k}} := \int_{\mathbb{T}^d} f(\mathbf{x}) \exp(-2\pi i \mathbf{k} \cdot \mathbf{x}) d\mathbf{x}$, $\mathbf{k} \in \mathbb{Z}^d$, of f , decay like $|\hat{f}_{\mathbf{k}}| \sim \prod_{i=1}^d (1 + |k_i|^2)^{-\alpha/2}$ for $\alpha \in \{5/4, 2, 6\}$ and, consequently, $f \in H_{\text{mix}}^s(\mathbb{T}^d)$ with $s = \alpha - 1/2 - \varepsilon$ for $\varepsilon > 0$.

Test function f from $H_{\text{mix}}^{3/4-\varepsilon}(\mathbb{T}^d)$

We start with the test function

$$f: \mathbb{T}^d \rightarrow \mathbb{R}, \quad f(\mathbf{x}) := \left(\frac{3}{2}\right)^{d/2} \prod_{i=1}^d \left(1 - |2(x_i \bmod 1) - 1|\right)^{1/4},$$

where we have for the Fourier coefficients $|\hat{f}_{\mathbf{k}}| \sim (\prod_{i=1}^d (1 + |k_i|^2)^{1/2})^{-5/4}$ and, consequently, $f \in H_{\text{mix}}^{3/4-\varepsilon}(\mathbb{T}^d)$, $\varepsilon > 0$.

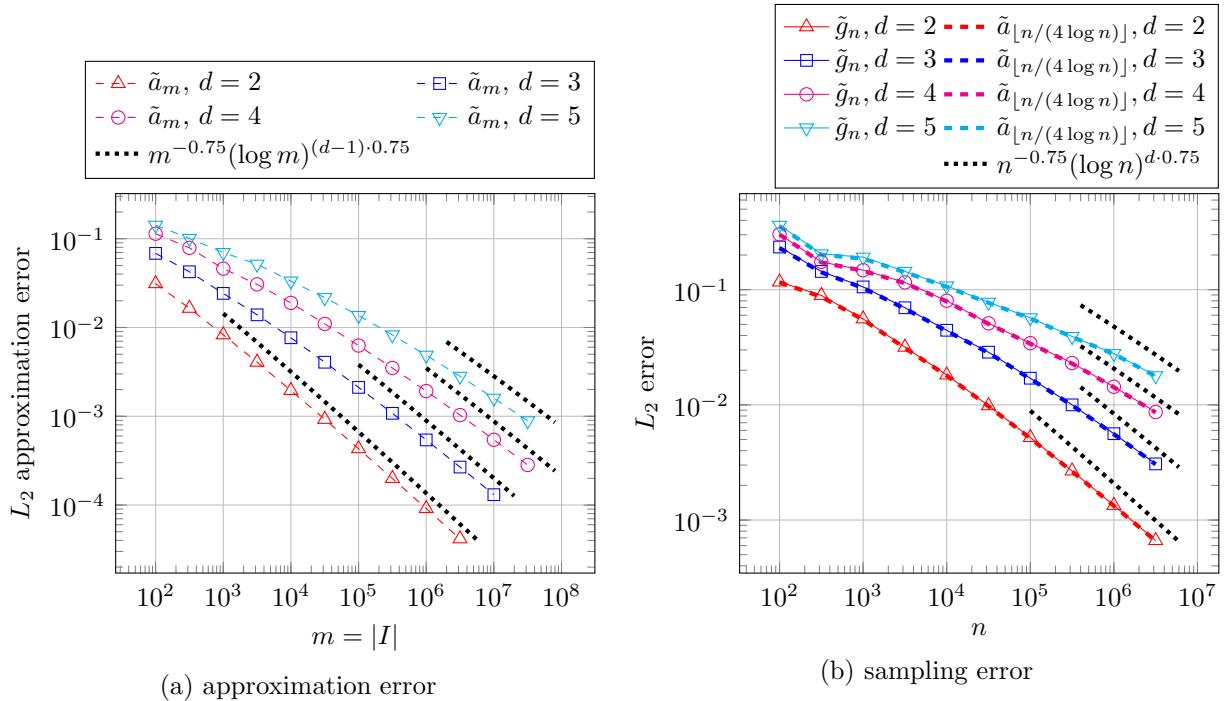


Figure 3: Approximation errors and least squares sampling errors for test function $f \in H_{\text{mix}}^{3/4-\varepsilon}(\mathbb{T}^d)$.

In Figure 3a, we visualize the relative approximation errors $\tilde{a}_m := \|f - P_{m-1}f\|_{L_2(\mathbb{T}^d)}$ for spatial dimensions $d = 2, 3, 4, 5$. Due to (8.6), these errors should decay like $m^{-0.75+\varepsilon}(\log m)^{(d-1)\cdot(0.75-\varepsilon)}$ for sufficiently large m . Correspondingly, we plot $m^{-0.75}(\log m)^{(d-1)\cdot 0.75}$ as black dotted graphs. We observe that the obtained approximation errors nearly decay as the theory suggests.

Next, we apply Algorithm 1 on the test function f using n randomly selected sampling nodes as sampling scheme. We do not compute the least squares solution directly

but use the iterative method LSQR [52] on the matrix \mathbf{L}_m , $m = \lfloor n/(4 \log n) \rfloor$. The obtained sampling errors $\tilde{g}_n := \|f - S_{\mathbf{X}}^m f\|_{L_2(\mathbb{T}^d)}$ are visualized in Figure 3b as well as the graphs $\sim n^{-0.75}(\log n)^{d-0.75}$ as dotted lines which correspond to the theoretical upper bounds $n^{-0.75+\varepsilon}(\log n)^{d(0.75-\varepsilon)}$, cf. (8.7). We set the tolerance parameter of LSQR to $5 \cdot 10^{-8}$ and the maximum number of iterations to 100. For $d = 2$ and $d = 3$, the errors nearly decay like these bounds. For $d = 4$ and $d = 5$, the errors seem to decay slightly slower than the bounds. In order to investigate this further, we also plot the corresponding approximation errors \tilde{a}_m with $m = \lfloor n/(4 \log n) \rfloor$ as thick dashed lines. We observe that these approximation errors \tilde{a}_m , which are the best possible errors that can be achieved in this setting, almost coincide with the sampling errors. This means that we might still observe preasymptotic behavior.

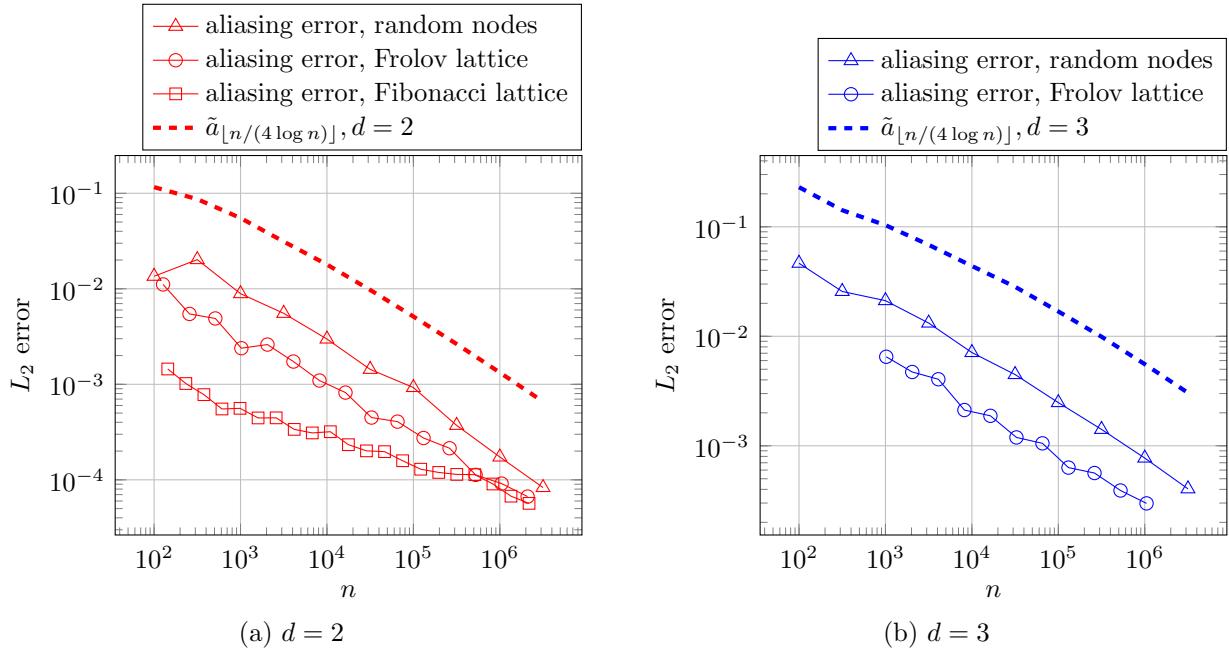


Figure 4: Least squares aliasing errors and approximation errors for test function $f \in H_{\text{mix}}^{3/4-\varepsilon}(\mathbb{T}^d)$.

For $d = 2$ spatial dimensions, we have a closer look at the sampling errors. In Figure 4a, we again plot the approximation errors \tilde{a}_m , $m = \lfloor n/(4 \log n) \rfloor$. In addition, the aliasing errors $\|P_{m-1}f - S_{\mathbf{X}}^m f\|_{L_2(\mathbb{T}^d)}$, $m = \lfloor n/(4 \log n) \rfloor$, which are the errors caused by Algorithm 1 since

$$\|f - S_{\mathbf{X}}^m f\|_{L_2(\mathbb{T}^d)}^2 = \|f - P_{m-1}f\|_{L_2(\mathbb{T}^d)}^2 + \|P_{m-1}f - S_{\mathbf{X}}^m f\|_{L_2(\mathbb{T}^d)}^2,$$

are shown as triangles. We observe that the aliasing errors nearly decay like the approximation errors, and that they are by one order of magnitude smaller. This corresponds to the behavior we observed in Figure 3b.

Moreover, we compare least squares using random point sets with least squares using quasi-random point sets. In particular, so-called Frolov lattices [29]¹ are considered as sampling sets in $d = 2$ spatial dimensions and used in Algorithm 1. The resulting sampling errors almost coincide with the approximation errors. The aliasing errors are visualized in Figure 4a as circles. It is remarkable that they decay similarly and are even lower than

¹In our numerical tests, we used the Frolov lattices constructed by the methods presented in [29] which has been published at <https://ins.uni-bonn.de/content/software-frolov>.

the aliasing errors for random nodes in most cases. A similar behavior can be observed for dimension $d = 3$, cf. Figure 4b.

In addition, we consider Fibonacci lattices in dimension $d = 2$, cf. Figure 4a. For $n \geq 832040$, the matrices \mathbf{L}_m , $m = \lfloor n/(4 \log n) \rfloor$, contain at least two identical columns and correspondingly, the smallest eigenvalue of $\mathbf{L}_m^* \mathbf{L}_m$ is zero. Therefore, obtaining $S_{\mathbf{X}}^m f$ via Algorithm 1 is not possible if the least squares solution is computed directly. An iterative method like LSQR may still work but the number of iterations may have to be restricted. In Figure 4a, the obtained aliasing errors via LSQR are shown as squares, and they are smaller than in the other cases but they seem to decay slower. However, when we decreased the tolerance parameter of the LSQR algorithm, we observed for $n \geq 832040$ aliasing errors and sampling errors larger than 1.

Kink test function f from $H_{\text{mix}}^{3/2-\varepsilon}(\mathbb{T}^d)$

Next, we consider the kink test function $f: \mathbb{T}^d \rightarrow \mathbb{R}$,

$$f(\mathbf{x}) = \prod_{i=1}^d \left(\frac{15}{4\sqrt{3}} \cdot 5^{3/4} \cdot \max \left(\frac{1}{5} - \left((x_i \bmod 1) - \frac{1}{2} \right)^2, 0 \right) \right) \in H_{\text{mix}}^{3/2-\varepsilon}(\mathbb{T}^d)$$

with Fourier coefficients

$$\hat{f}_{\mathbf{k}} = \prod_{i=1}^d \begin{cases} \frac{5^{5/4}\sqrt{3}}{8} (-1)^{k_i} \frac{\sqrt{5} \sin(2k_i\pi/\sqrt{5}) - 2k_i\pi \cos(2k_i\pi/\sqrt{5})}{\pi^3 k_i^3} & \text{for } k_i \neq 0, \\ \frac{5^{1/4}}{\sqrt{3}} & \text{for } k_i = 0. \end{cases}$$

Besides the different test function f , we use the same setting as before.

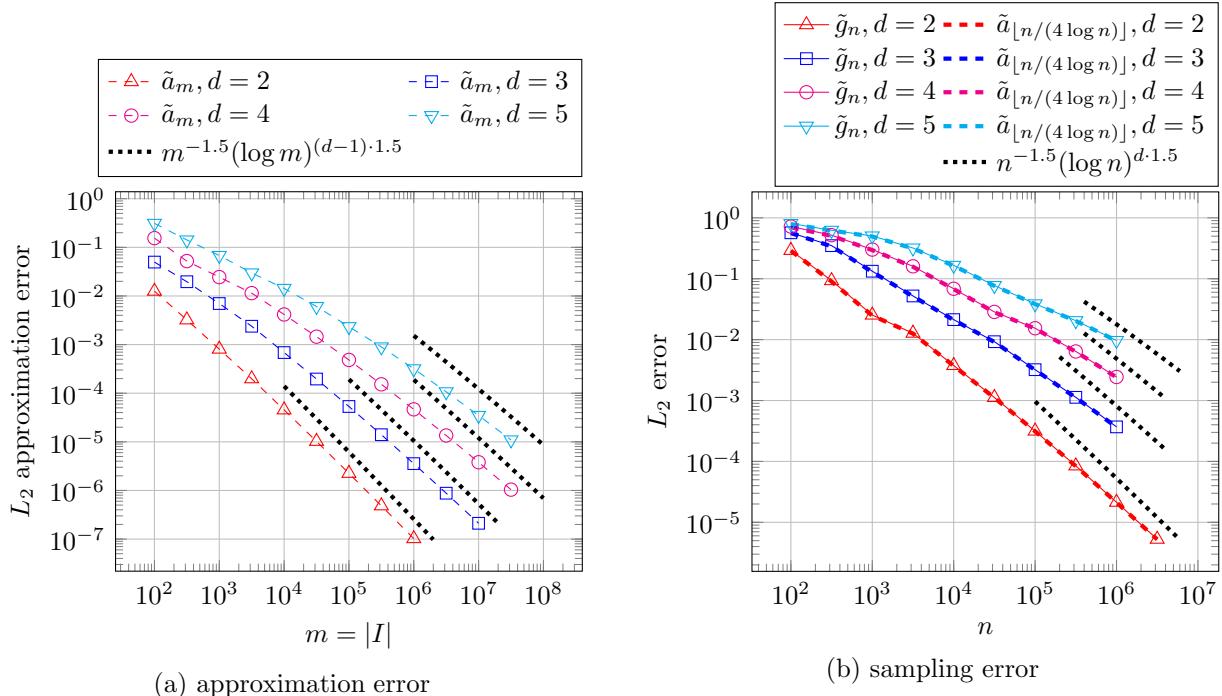


Figure 5: Approximation errors and least squares sampling errors for test function $f \in H_{\text{mix}}^{3/2-\varepsilon}(\mathbb{T}^d)$.

In Figure 5a, we visualize the relative approximation errors $\tilde{a}_m := \|f - P_{m-1}f\|_{L_2(\mathbb{T}^d)}$ for spatial dimensions $d = 2, 3, 4, 5$. These errors should decay like $m^{-1.5+\varepsilon}(\log m)^{(d-1)\cdot(1.5-\varepsilon)}$ for sufficiently large m , and we observe that the obtained approximation errors nearly decay as the theoretical results suggest. Next, we apply Algorithm 1 with random nodes to the test function f using the iterative method LSQR. The resulting sampling errors $\tilde{g}_n := \|f - S_{\mathbf{X}}^m f\|_{L_2(\mathbb{T}^d)}$ are depicted in Figure 5b. In addition, the graphs $\sim n^{-1.5}(\log n)^{d-1.5}$ are shown as dotted lines which roughly correspond to the theoretical upper bounds $n^{-1.5+\varepsilon}(\log n)^{d\cdot(1.5-\varepsilon)}$. The errors seem to decay according to this bound for $d = 2$ and slower for $d = 3, 4, 5$. Again, we also plot the corresponding approximation errors \tilde{a}_m with $m = \lfloor n/(4 \log n) \rfloor$ as thick dashed lines, and we observe that these approximation errors $\tilde{a}_{\lfloor n/(4 \log n) \rfloor}$ almost coincide with the sampling errors. Correspondingly, we still have preasymptotic behavior for $d = 3, 4, 5$.

Test function f from $H_{\text{mix}}^{11/2-\varepsilon}(\mathbb{T}^d)$

As a third test function f , we consider an L_2 -normalized product of periodic one-dimensional B-Splines of order 6, where each factor depends on a single variable and is a piecewise polynomial of degree 5, and therefore, we have $f \in H_{\text{mix}}^{11/2-\varepsilon}(\mathbb{T}^d)$.

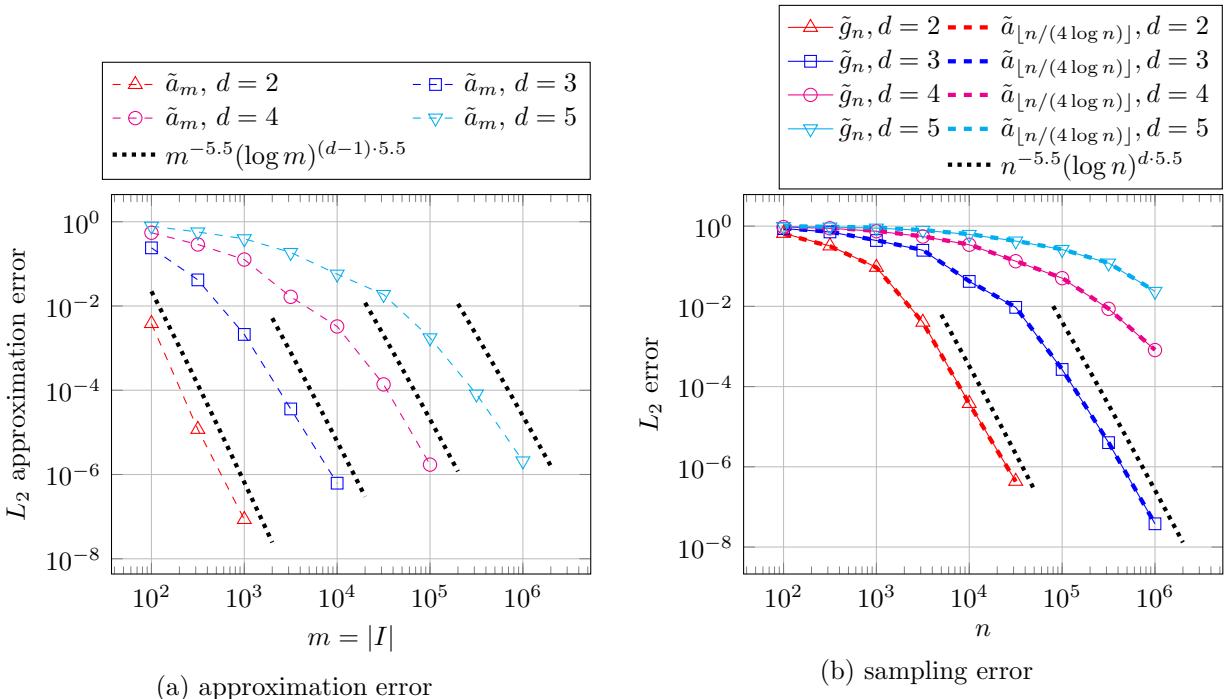


Figure 6: Approximation errors and least squares sampling errors for test function $f \in H_{\text{mix}}^{11/2-\varepsilon}(\mathbb{T}^d)$.

In Figure 6a, the relative approximation errors $\tilde{a}_m := \|f - P_{m-1}f\|_{L_2(\mathbb{T}^d)}$ are visualized for spatial dimensions $d = 2, 3, 4, 5$, which roughly decay like $m^{-5.5+\varepsilon}(\log m)^{(d-1)\cdot(5.5-\varepsilon)}$ for sufficiently large m . The sampling errors $\tilde{g}_n := \|f - S_{\mathbf{X}}^m f\|_{L_2(\mathbb{T}^d)}$ when applying Algorithm 1 with random nodes to the test function f using the iterative method LSQR are depicted in Figure 6b. In addition, the graphs $\sim n^{-5.5}(\log n)^{d-5.5}$ are plotted as dotted lines which correspond to the theoretical upper bounds. The errors seem to decay roughly according to this bound for $d = 2$ and $d = 3$ or slightly slower. Again, the corresponding approximation

errors \tilde{a}_m with $m = n/(4 \log n)$ are shown as thick dashed lines, and we observe that these approximation errors $\tilde{a}_{\lfloor n/(4 \log n) \rfloor}$ almost coincide with the sampling errors. This means we still have preasymptotic behavior.

10.2 Integration of functions from spaces with mixed smoothness

Extensive numerical tests on integration using random point sets were performed by Oettershagen [50], cf. in particular [50, Sec. 4.1] for tests concerning numerical integration in Sobolev spaces with mixed smoothness $H_{\text{mix}}^s(\mathbb{T}^d)$. These numerical tests, performed for $d \in \{2, 4, 8, 16\}$ spatial dimensions and smoothness $s \in \{1, 2, 3\}$, suggest that the worst case cubature error may decay with a main rate of n^{-s} with additional log factors. This is a remarkable behavior since plain Monte-Carlo with random points usually leads to an error decay of $n^{-1/2}$.

However, the corresponding theoretical results in [50, Sec. 4.2] only give a main rate of $n^{-s+1/2}$. We highlight that our results obtained in this paper bridge this gap of $1/2$ in the main rate, since we show a worst-case error of $\sim n^{-s}(\log n)^{ds}$ in Corollary 8.6, i.e., our theoretical main rate corresponds to the observations by Oettershagen [50]. Moreover, Algorithm 1 guarantees suitable error bounds in the preasymptotic setting, cf. Corollary 8.6 together with (8.5). More details on cubature rules based on least squares additionally refined using some variance reduction technique can be found in [42].

10.3 Recovery and integration for the non-periodic case

Now we consider the non-periodic situation. We use the Chebyshev measure $\tilde{\varrho}_D(d\mathbf{x}) := \prod_{t=1}^d (\pi \sqrt{1 - x_t^2})^{-1} d\mathbf{x}$ as random sampling scheme on $D = [-1, 1]^d$. We further define the non-periodic space $\tilde{H}_{\text{mix}}^s([-1, 1]^d)$ via the reproducing kernel

$$\tilde{K}_{s,*}^1(x, y) := 1 + 2 \sum_{h \in \mathbb{N}_0} \frac{\cos(h \arccos(x)) \cos(h \arccos(y))}{w_{s,*}(h)^2}, \quad x, y \in \mathbb{T},$$

and its tensor product

$$\tilde{K}_{s,*}^d(\mathbf{x}, \mathbf{y}) := K_{s,*}^1(x_1, y_1) \otimes \cdots \otimes K_{s,*}^1(x_d, y_d), \quad \mathbf{x}, \mathbf{y} \in \mathbb{T}^d.$$

The space $\tilde{H}_{\text{mix}}^s([-1, 1]^d)$ is embedded into $L_2(D, \tilde{\varrho}_D)$ if and only if $s > 1/2$. We denote the k -th basis index of η_k , $k = 1, \dots, m-1$, by $\mathbf{h}_k := (h_{k,t})_{t=1}^d \in \mathbb{N}_0^d$, and we have $\eta_1 \equiv 1$ and

$$\eta_k(\mathbf{x}) = \prod_{t=1}^d \sqrt{2}^{\min\{h_{k,t}, 1\}} \cos(h_{k,t} \arccos x_t)$$

if $k > 1$. Moreover, for the vector $\mathbf{b} \in \mathbb{C}^{m-1}$ in (7.1), we have

$$b_k := \int_D \eta_k d\mu_D = \prod_{t=1}^d \begin{cases} 2 & \text{for } h_{k,t} = 0, \\ \frac{2\sqrt{2}}{1-h_{k,t}^2} & \text{for } h_{k,t} \in 2\mathbb{N}, \\ 0 & \text{for } h_{k,t} \in 2\mathbb{N}-1, \end{cases} \quad (10.1)$$

for $\mu_D \equiv 1$.

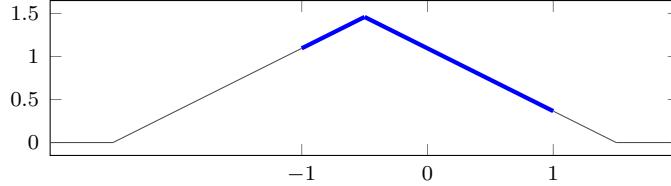


Figure 7: Dilated, scaled, and shifted B-Spline of order 2 considered in interval $[-1, 1]$.

For the numerical experiments we consider the test function

$$f: \mathbb{R} \rightarrow \mathbb{R}, \quad f(\mathbf{x}) := \left(\frac{3\pi}{49\pi - 48\sqrt{3}} \right)^{d/2} \prod_{i=1}^d \begin{cases} 5 + 2x & \text{for } -5/2 \leq x_i < -1, \\ 3 - 2x & \text{for } -1 \leq x_i < 3/2, \\ 0 & \text{otherwise,} \end{cases}$$

on $D := [-1, 1]^d$, where the one-dimensional version is depicted in Figure 7. For the Chebyshev coefficients of f , we have

$$\hat{f}_{\mathbf{k}} = \left(\frac{3\pi}{49\pi - 48\sqrt{3}} \right)^{d/2} \prod_{i=1}^d \begin{cases} 4 \cdot \frac{\sqrt{3} k_i \cos(2\pi k_i / 3) + \sin(2\pi k_i / 3)}{\sqrt{2}(-k_i + k_i^3)\pi} & \text{for } k_i \geq 2, \\ -(3\sqrt{6} + 2\sqrt{2}\pi)/(6\pi) & \text{for } k_i = 1, \\ 11/3 - 2\sqrt{3}/\pi & \text{for } k_i = 0, \end{cases}$$

and consequently, $f \in \tilde{H}_{\text{mix}}^{3/2-\varepsilon}([-1, 1]^d)$ for $\varepsilon > 0$.

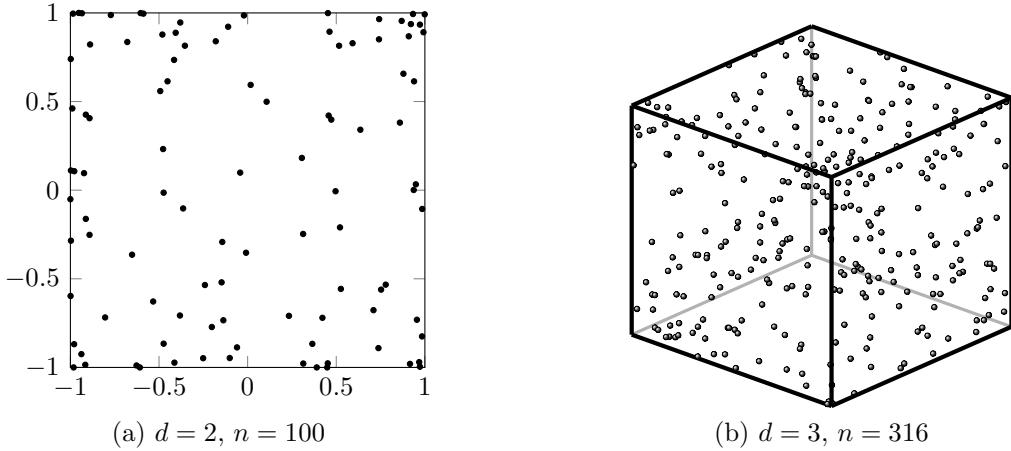


Figure 8: Realizations of random nodes with respect to the Chebyshev measure $\tilde{\varrho}_D(\mathbf{x})$.

We use the parameters as in Section 10.1 and apply Algorithm 1 on the test function f in the non-periodic setting, where we generate the random nodes with respect to the measure $\tilde{\varrho}_D(d\mathbf{x}) := \prod_{i=1}^d (\pi \sqrt{1 - x_i^2})^{-1} dx_i$. In Figure 8, we show realizations for these random nodes. As before, we do not compute the least squares solution directly but use the iterative method LSQR on the matrix \mathbf{L}_m , $m = \lfloor n/(4 \log n) \rfloor$. The obtained sampling errors $\tilde{g}_n := \|f - S_{\mathbf{X}}^m f\|_{L_2([-1, 1]^d, \tilde{\varrho}_D)}$ with $m = \lfloor n/(4 \log n) \rfloor$ are plotted in Figure 9a as triangles as well as the corresponding approximation errors $\tilde{a}_m := \|f - P_{m-1} f\|_{L_2([-1, 1]^d, \tilde{\varrho}_D)}$ as thick dashed lines. We observe that the sampling and approximation errors almost coincide. Moreover, we plot the graphs $\sim n^{-1.5}(\log n)^{d+1.5}$ as dotted lines which correspond to the expected theoretical

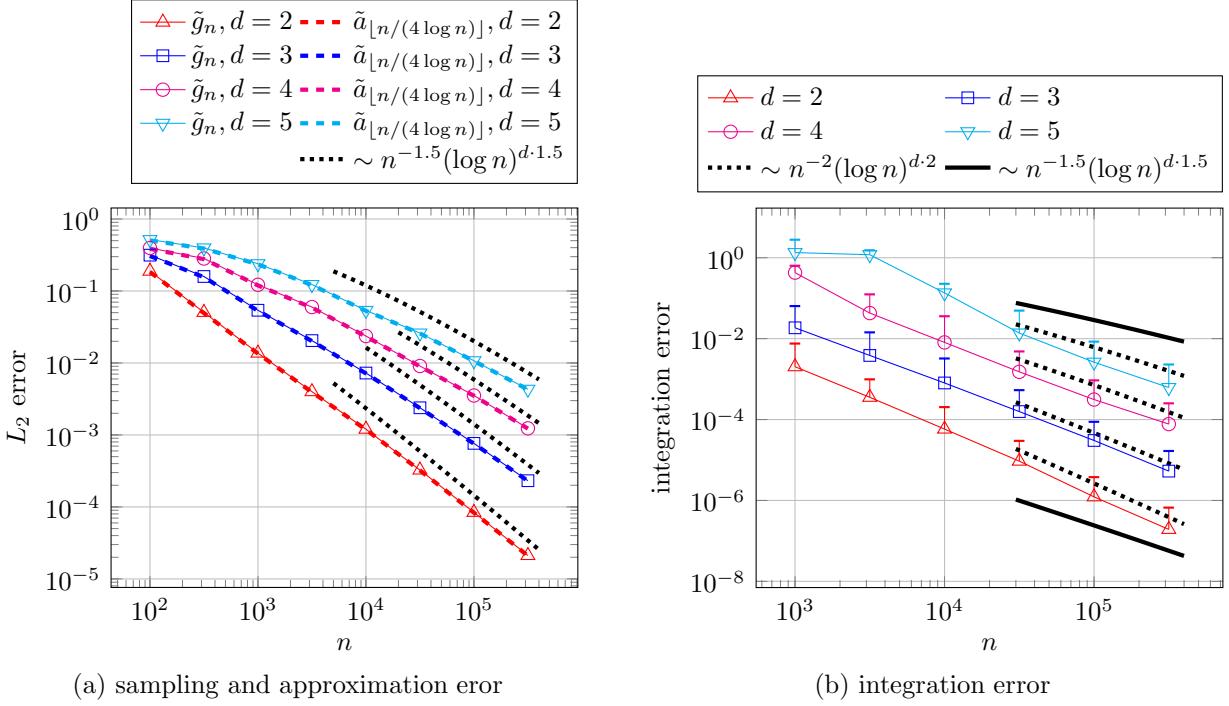


Figure 9: Sampling errors and integration errors for non-periodic test function $f \in \hat{H}_{\text{mix}}^{3/2-\varepsilon}([-1, 1]^d)$.

upper bounds $n^{-1.5+\varepsilon}(\log n)^{d+(1.5-\varepsilon)}$. We observe that the obtained numerical errors nearly decay like these theoretical upper bound.

In addition, we use the numerically computed Chebyshev coefficients c_k from Algorithm 1 to compute the approximation $Q_{\mathbf{X}}f$ of $I(f)$ by $Q_{\mathbf{X}}f = \int_D S_{\mathbf{X}}^m f \, d\mu_D = \sum_{k=1}^m c_k b_k$ where the complex numbers b_k are calculated as stated in (10.1). We repeatedly perform each test 100 times with different random nodes. The averages for the integration errors $|I(f) - Q_{\mathbf{X}}f|$ of the 100 test runs are depicted as triangles in Figure 9b and the maxima as error bars. Moreover, we plot the graphs $\sim n^{-2}(\log n)^{d+2}$ as dotted lines, and we observe that the obtained integration errors approximately decay like these graphs. For comparison, we also plot $n^{-1.5}(\log n)^{d+1.5}$ for $d = 2$ and $d = 5$ as thick solid lines which belong to the theoretical results $n^{-1.5+\varepsilon}(\log n)^{d+(1.5-\varepsilon)}$ one obtains analogously to (8.9) in Section 8 for the non-periodic case. These thick solid lines decay distinctly slower.

In particular, we strongly expect that the theoretical preasymptotic results in (8.5) and [32] also hold for the Chebyshev case.

Acknowledgment. The authors would like to thank Bastian Bohn, Michael Griebel, Karl-heinz Gröchenig, Aicke Hinrichs, Boris S. Kashin, David Krieg, Thomas Kühn, Laura Lipper, Erich Novak, Jens Oettershagen, Peter Oswald, Daniel Potts, Holger Rauhut, Ingo Steinwart, Martin Stoll, Vladimir N. Temlyakov and Mario Ullrich for several discussions on the topic. L.K. gratefully acknowledges the funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, project number 380648269). T.U. would like to acknowledge support by the DFG Ul-403/2-1. T.V. gratefully acknowledges support by the SAB 100378180.

References

- [1] R. F. Bass and K. Gröchenig. Relevant sampling of band-limited functions. *Illinois J. Math.*, 57(1):43–58, 2013.
- [2] A. Belhadji, R. Bardenet, and P. Chainais. Kernel quadrature with DPPs. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- [3] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic Publishers, Boston, MA, 2004. With a preface by Persi Diaconis.
- [4] C. Bernardi and Y. Maday. Polynomial interpolation results in Sobolev spaces. *J. Comput. Appl. Math.*, 43(1):53–80, 1992.
- [5] B. Bohn. *Error analysis of regularized and unregularized least-squares regression on discretized function spaces*. Dissertation, Institut für Numerische Simulation, Universität Bonn, 2017.
- [6] B. Bohn. On the convergence rate of sparse grid least squares regression. In J. Garcke, D. Pflüger, C. Webster, and G. Zhang, editors, *Sparse Grids and Applications - Miami 2016*, volume 123 of *Lecture Notes in Computational Science and Engineering*, pages 19–41. Springer, 2018. Also available as INS Preprint no 1711.
- [7] B. Bohn and M. Griebel. Error estimates for multivariate regression on discretized function spaces. *SIAM J. Numer. Anal.*, 55(4):1843–1866, 2017.
- [8] G. Byrenheid. *Sparse representation of multivariate functions based on discrete point evaluations*. Dissertation, Institut für Numerische Simulation, Universität Bonn, 2018.
- [9] G. Byrenheid, L. Kämmerer, T. Ullrich, and T. Volkmer. Tight error bounds for rank-1 lattice sampling in spaces of hybrid mixed smoothness. *Numer. Math.*, 136(4):993–1034, 2017.
- [10] A. Chkifa, A. Cohen, G. Migliorati, F. Nobile, and R. Tempone. Discrete least squares polynomial approximation with random evaluations – application to parametric and stochastic elliptic PDEs. *ESAIM: M2AN*, 49(3):815–837, 2015.
- [11] A. Christmann and I. Steinwart. *Support Vector Machines*. Springer, 2008.
- [12] F. Cobos, T. Kühn, and W. Sickel. Optimal approximation of multivariate periodic Sobolev functions in the sup-norm. *J. Funct. Anal.*, 270(11):4196–4212, 2016.
- [13] A. Cohen, M. A. Davenport, and D. Leviatan. On the stability and accuracy of least squares approximations. *Found. Comput. Math.*, 13(5):819–834, 2013.
- [14] A. Cohen, M. A. Davenport, and D. Leviatan. Correction to: On the stability and accuracy of least squares approximations [MR3105946]. *Found. Comput. Math.*, 19(1):239, 2019.
- [15] A. Cohen and G. Migliorati. Optimal weighted least-squares methods. *SMAI J. Comput. Math.*, 3:181–203, 2017.
- [16] D. Dün. B-spline quasi-interpolant representations and sampling recovery of functions with mixed smoothness. *J. Complexity*, 27(6):541–567, 2011.
- [17] D. Dün, V. N. Temlyakov, and T. Ullrich. *Hyperbolic Cross Approximation*. Advanced Courses in Mathematics. CRM Barcelona. Birkhäuser/Springer, 2019.
- [18] D. Dün and T. Ullrich. Lower bounds for the integration error for multivariate functions with mixed smoothness and optimal Fibonacci cubature for functions on the square. *Math. Nachr.*, 288(7):743–762, 2015.
- [19] F. Daĭ, A. Primak, V. N. Temlyakov, and S. Y. Tikhonov. Integral norm discretization and related problems. *Uspekhi Mat. Nauk*, 74(4(448)):3–58, 2019.
- [20] R. A. DeVore, S. V. Konyagin, and V. N. Temlyakov. Hyperbolic wavelet approximation. *Constr. Approx.*, 14(1):1–26, 1998.
- [21] J. Dick, F. Y. Kuo, and I. H. Sloan. High-dimensional integration: the quasi-Monte Carlo way. *Acta Numer.*, 22:133–288, 2013.
- [22] C. F. Dunkl and Y. Xu. *Orthogonal polynomials of several variables*. Cambridge University Press, 1. edition, 2001.
- [23] S. Foucart and H. Rauhut. *A mathematical introduction to compressive sensing*. Applied and Numerical Harmonic Analysis. Birkhäuser/Springer, New York, 2013.
- [24] K. Gröchenig. Irregular sampling, Toeplitz matrices, and the approximation of entire functions of exponential type. *Math. Comp.*, 68(226):749–765, 1999.
- [25] K. Gröchenig. Sampling, Marcinkiewicz-Zygmund inequalities, approximation, and quadrature rules. *J. Approx. Theory*, 257:105455, 2020.

- [26] A. Hinrichs. Optimal importance sampling for the approximation of integrals. *J. Complexity*, 26(2):125–134, 2010.
- [27] A. Hinrichs, E. Novak, D. Krieg, J. Prochno, and M. Ullrich. On the power of random information. In F. J. Hickernell and P. Kritzer, editors, *Multivariate Algorithms and Information-Based Complexity*, volume 27 of *Radon Series on Computational and Applied Mathematics*, pages 43–64. De Gruyter, Berlin/Boston, 2020.
- [28] A. Hinrichs, E. Novak, and J. Vybíral. Linear information versus function evaluations for L_2 -approximation. *J. Approx. Theory*, 153(1):97–107, 2008.
- [29] C. Kacwin, J. Oettershagen, M. Ullrich, and T. Ullrich. Numerical performance of optimized Frolov lattices in tensor product reproducing kernel Sobolev spaces. *Found. Comput. Math.*, 2020. <https://doi.org/10.1007/s10208-020-09463-y>.
- [30] L. Kämmerer, D. Potts, and T. Volkmer. Approximation of multivariate periodic functions by trigonometric polynomials based on rank-1 lattice sampling. *J. Complexity*, 31(4):543–576, 2015.
- [31] L. Kämmerer, D. Potts, and T. Volkmer. Approximation of multivariate periodic functions by trigonometric polynomials based on sampling along rank-1 lattice with generating vector of Korobov form. *J. Complexity*, 31(3):424–456, 2015.
- [32] D. Krieg. Tensor power sequences and the approximation of tensor product operators. *J. Complexity*, 44:30–51, 2018.
- [33] D. Krieg and M. Ullrich. Function values are enough for L_2 -approximation. *Found. Comput. Math.*, 2020. <https://doi.org/10.1007/s10208-020-09481-w>.
- [34] T. Kühn. New preasymptotic estimates for the approximation of periodic Sobolev functions. In J. de Gier, C. E. Praeger, and T. Tao, editors, *2018 MATRIX annals*, volume 3 of *MATRIX Book Ser.*, pages 97–112. Springer International Publishing, 2020.
- [35] T. Kühn, W. Sickel, and T. Ullrich. Approximation of mixed order Sobolev functions on the d -torus: asymptotics, preasymptotics and d -dependence. *Constr. Approx.*, 42(3):353–398, 2015.
- [36] T. Kühn, W. Sickel, and T. Ullrich. How anisotropic mixed smoothness affects the decay of singular numbers of Sobolev embeddings. *J. Complexity*, 63:101523, 2021.
- [37] F. Y. Kuo, G. W. Wasilkowski, and H. Woźniakowski. On the power of standard information for multivariate approximation in the worst case setting. *J. Approx. Theory*, 158(1):97–125, 2009.
- [38] W. Lu and H. Wang. On the power of standard information for tractability for L_2 -approximation in the average case setting. *arXiv:2101.05200*, 2021.
- [39] W. Lu and H. Wang. On the power of standard information for tractability for L_2 -approximation in the randomized setting. *arXiv:2101.03665*, 2021.
- [40] F. Lust-Piquard and G. Pisier. Noncommutative Khintchine and Paley inequalities. *Ark. Mat.*, 29(2):241–260, 1991.
- [41] A. W. Marcus, D. A. Spielman, and N. Srivastava. Interlacing families II: Mixed characteristic polynomials and the Kadison-Singer problem. *Ann. of Math. (2)*, 182(1):327–350, 2015.
- [42] G. Migliorati and F. Nobile. Stable high-order randomized cubature formulae in arbitrary dimension. *arXiv:1812.07761v4*, 2020.
- [43] M. Moeller and T. Ullrich. L_2 -norm sampling discretization and recovery of functions from RKHS with finite trace. *arXiv:2009.11940*, 2020.
- [44] N. Nagel, M. Schäfer, and T. Ullrich. A new upper bound for sampling numbers. *Found. Comput. Math.*, to appear.
- [45] D. Needell, N. Srebro, and R. Ward. Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm. *Math. Program.*, 155(1-2, Ser. A):549–573, 2016.
- [46] E. Novak and H. Woźniakowski. *Tractability of multivariate problems. Vol. 1: Linear information*, volume 6 of *EMS Tracts in Mathematics*. European Mathematical Society (EMS), Zürich, 2008.
- [47] E. Novak and H. Woźniakowski. *Tractability of multivariate problems. Volume II: Standard information for functionals*, volume 12 of *EMS Tracts in Mathematics*. European Mathematical Society (EMS), Zürich, 2010.
- [48] E. Novak and H. Woźniakowski. On the power of function values for the approximation problem in various settings. *Surv. Approx. Theory*, 6:1–23, 2011.
- [49] E. Novak and H. Woźniakowski. *Tractability of multivariate problems. Volume III: Standard information for operators*, volume 18 of *EMS Tracts in Mathematics*. European Mathematical Society (EMS), Zürich, 2012.
- [50] J. Oettershagen. *Construction of Optimal Cubature Algorithms with Applications to Econometrics and*

- Uncertainty Quantification.* Dissertation, Institut für Numerische Simulation, Universität Bonn, 2017.
- [51] R. I. Oliveira. Sums of random Hermitian matrices and an inequality by Rudelson. *Electr. Comm. Probab.*, 15:203–212, 2010.
 - [52] C. C. Paige and M. A. Saunders. LSQR: An algorithm for sparse linear equations and sparse least squares. *ACM Trans. Math. Software*, 8(1):43–71, 1982.
 - [53] A. Papageorgiou and H. Woźniakowski. Tractability through increasing smoothness. *J. Complexity*, 26(5):409–421, 2010.
 - [54] H. Rauhut. Compressive sensing and structured random matrices. In M. Fornasier, editor, *Theoretical Foundations and Numerical Methods for Sparse Recovery*, volume 9 of *Radon Series on Computational and Applied Mathematics*. De Gruyter, Berlin, 2010.
 - [55] H. Rauhut and R. Ward. Sparse Legendre expansions via ℓ_1 -minimization. *J. Approx. Theory*, 164(5):517–533, 2012.
 - [56] H. Rauhut and R. Ward. Interpolation via weighted ℓ_1 minimization. *Appl. Comput. Harmon. Anal.*, 40(2):321–351, 2016.
 - [57] M. Rudelson. Random vectors in the isotropic position. *J. Funct. Anal.*, 164(1):60–72, 1999.
 - [58] W. Sickel and T. Ullrich. Tensor products of Sobolev-Besov spaces and applications to approximation from the hyperbolic cross. *J. Approx. Theory*, 161(2):748–786, 2009.
 - [59] S. Smale and D.-X. Zhou. Shannon sampling and function reconstruction from point values. *Bull. Amer. Math. Soc. (N.S.)*, 41(3):279–305, 2004.
 - [60] V. N. Temlyakov. *Approximation of periodic functions*. Computational Mathematics and Analysis Series. Nova Science Publishers Inc., Commack, NY, 1993.
 - [61] V. N. Temlyakov. The Marcinkiewicz-type discretization theorems. *Constr. Approx.*, 48(2):337–369, 2018.
 - [62] V. N. Temlyakov. Sampling discretization error of integral norms for function classes. *J. Complexity*, 54:101408, 2019.
 - [63] V. N. Temlyakov. On optimal recovery in L_2 . *J. Complexity*, to appear.
 - [64] M. Ullrich. On the worst-case error of least squares algorithms for L_2 -approximation with high probability. *Journal of Complexity*, 60, 2020.
 - [65] G. W. Wasilkowski and H. Woźniakowski. On the power of standard information for weighted approximation. *Found. Comput. Math.*, 1(4):417–434, 2001.
 - [66] G. W. Wasilkowski and H. Woźniakowski. The power of standard information for multivariate approximation in the randomized setting. *Math. Comp.*, 76(258):965–988, 2007.