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

March 23, 2020

Abstract

We consider a least squares regression algorithm for the recovery of complex-valued functions belonging to a reproducing kernel Hilbert space $H(K)$ from random data measuring the error in $L_2(D, \varrho_D)$. We prove worst-case recovery guarantees and improve on the recent new upper bounds for sampling numbers in Krieg, M. Ullrich [32] by explicitly controlling all the involved constants with respect to the underlying spatial dimension d. This leads to new preasymptotic recovery bounds with high probability for the error of *Hyperbolic Fourier Regression* for multivariate functions. In addition, we analyze the algorithm *Hyperbolic Wavelet Regression* also based on least-squares to recover non-periodic functions from random samples. As a further application we reconsider the analysis of a cubature method based on plain random points with optimal weights introduced by Oettershagen in [45]. We confirm a conjecture (which was based on various numerical experiments in [45]) and give improved 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. Last but not least, we contribute new preasymptotic bounds to the problem of the recovery of individual functions from n samples which has been already considered by Smale, Zhou [55], Bohn, Griebel [6], Cohen, Davenport, Leviatan [12], Chkifa, Migliorati, Nobile, Tempone [9], Cohen, Migliorati [14], Krieg [31] and several others.

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 a complex-valued multivariate function on a domain $D \subset \mathbb{R}^d$ from function samples on the nodes $\mathbf{X} = (\mathbf{x}^1, ..., \mathbf{x}^n) \in D^n$, which are drawn independently according to a probability measure ϱ_D. This may be interpreted as reconstructing a function from so-called scattered data, see [60]. The functions are modeled as elements

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

from some reproducing kernel Hilbert space $H(K)$ with kernel $K(\cdot,\cdot)$, see [3], [10]. The error is measured in $L_2(D,\varrho_D)$. This problem has been considered by several authors in the literature, e.g. Smale, Zhou [55], Bohn [4, 5], Bohn, Griebel [6], Cohen, Davenport, Leviatan [12], Chkifa, Migliorati, Nobile, Tempone [9], Cohen, Migliorati [14], and many others to mention just a few. In these contributions the authors studied the problem of recovering an individual function from random observations.

Our main focus in this paper is the uniform recovery of functions belonging to a class $H(K)$ and to give worst-case recovery guarantees. In fact, the results below have a flavor similar to the use of the well-known "probabilistic method". This approach is rather popular in the Information Based Complexity (IBC) community, where it is also referred to as "random information", see the recent survey [25]. Note that this probabilistic method has been recently applied successfully in the field of "compressed sensing", see [21]. In fact, we aim at recovering all $f\in H(K)$ from sampled values at $\mathbf{X}=(\mathbf{x}^1,...,\mathbf{x}^n)$ simultaneously with high probability, where the error is measured in $L_2(D,\varrho_D)$. To be more precise, we provide bounds for the worst-case error

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

We construct a recovery operator $S^m_{\mathbf{X}}$ which computes a best least squares fit $S^m_{\mathbf{X}}f$ to the given data $(f(\mathbf{x}^1),...,f(\mathbf{x}^n))^\top$ from the finite-dimensional space spanned by the first m singular vectors of the embedding

$$\mathrm{Id}:H(K)\to L_2(D,\varrho_D)\,. \tag{1.1}$$

The right singular vectors $e_1(\cdot),e_2(\cdot),...$ of this embedding are arranged according to their importance, i.e., the non-increasing rearrangement of the singular values $\sigma_1\ge\sigma_2\ge\cdots>0$. As already mentioned, the nodes $\mathbf{X}$ are drawn at random from the domain D according to ϱ_D. However, in contrast to the Monte-Carlo setting we use the same reconstruction operator $S^m_{\mathbf{X}}$ for the entire class $H(K)$.

The investigations of this paper are inspired by the recent results by Krieg and M. Ullrich [32], see the discussion in Remark 5.8 and Section 8.3. In this paper we extend and improve the results from [32] 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 $\mathrm{Id}^*\circ\mathrm{Id}:H(K)\to H(K)$.

Theorem 1.1. *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 $K:D\times D\to\mathbb{C}$ satisfies $\sup_{\mathbf{x}\in D}K(\mathbf{x},\mathbf{x})<\infty$. Let further ϱ_D denote a probability measure on D. Furthermore, for $n\in\mathbb{N}$ we define $m\in\mathbb{N}$ such that*

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

holds. Then the random reconstruction operator $S^m_{\mathbf{X}}$ *(see Algorithm 1 below), which uses* n *samples on the i.i.d. according to* ϱ_D *drawn nodes* $\mathbf{X} = (\mathbf{x}^1, ..., \mathbf{x}^n)$*, satisfies for* $0 < \delta < 1/2$

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

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

The occurrence of the fundamental quantity $N(m)$, which is sometimes called spectral function (see [23] and the references therein), is certainly not a surprise. It represents a well-known ingredient for inequalities related to sampling and discretization, see for instance Gröchenig and Bass [1], Gröchenig [22, 23], Temlyakov [57, 58], and Temlyakov et al. [18]. In cases where $T(m) \sim n\,\sigma_m^2$ up to logarithmic factors we observe nearly optimal error bounds with respect to the first $m-1$ basis functions $e_k(\cdot)$ that are used to compute $S^m_{\mathbf{X}} f \in \operatorname{span}\{e_1, \ldots, e_{m-1}\}$. In addition, for $N(m) \sim m$ we also achieve near optimal error bounds with respect to the number of used sampling values n. Although the statement of Theorem 1.1 is highly general we may extract new bounds in several relevant situations. In this paper, we discuss *Hyperbolic Fourier Regression* and *Hyperbolic Wavelet Regression* in detail. Note that if the system of right singular vectors $(e_k)_{k\in\mathbb{N}}$ in Theorem 1.1 has the additional property that the orthonormal system $(\eta_k)_{k\in\mathbb{N}}$ (in $L_2(D, \varrho_D)$) given by $\eta_k := \sigma_k^{-1} e_k$ is uniformly bounded, i.e. $\|\eta_k\|_\infty \leq B$ for all k, then the bounds in Theorem 1.1 can be made much more precise, see Section 8.2 below. In fact, as a consequence we do not only see the bound in [32], we also have complete knowledge of all involved constants. Furthermore, our result is no longer restricted to the case of real-valued functions. In the proof we combine a modification of a recent concentration inequality for sums of Hermitian matrices due to Oliveira [46] with a symmetrization technique involving the complex version of Rudelson's lemma, see Lemma 4.3 below. The precise constants in this powerful inequality have been computed by Rauhut [49] based on a technique involving a non-commutative Khintchine inequality.

Note that in case of the complex Fourier system we have $B = 1$ and obtain dimension-free constants. We emphasize that within our worst-case error analysis we keep track of all the involved constants and give preasymptotic bounds where it is possible. 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 for *Hyperbolic Fourier Regression*, we discuss and apply recent preasymptotic bounds for the singular values $(\sigma_k)_{k\in\mathbb{N}}$, see [34, 35] and [30]. For the non-periodic situation we will use a modification of the algorithm called *Hyperbolic Wavelet Regression* where we use orthonormal compactly supported wavelets which are not uniformly bounded in L_∞. However, the quantities $N(m)$ and $T(m)$ can be controlled in this situation. In particular, the matrices appearing in the least squares problem are very sparse which gives an additional acceleration for the runtime of the algorithm.

In the above result we control the failure probability and sample with respect to the known density ϱ_D. One may implement and use this result in practice since the error guarantees are reliable with high probability, see Section 10. In addition, a technique called "optimal importance sampling", see e.g. [24, 50, 51, 39], turns out to be useful in this context, see Algorithm 2 below. As proposed in [14] and specified in full detail in [32], one may sample from a reweighted distribution/density ϱ_m, defined in (5.13) 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. A refinement of this technique together with our Theorem 1.1 above leads to the following precise bound on the sampling numbers $g_n(\mathrm{Id} : H(K) \to L_2(D, \varrho_D))$, defined in Section 5.3, under even weaker conditions compared to Theorem 1.1.

Theorem 1.2. *Let $H(K)$ be a separable reproducing kernel Hilbert space of complex-valued functions defined on $D \subset \mathbb{R}^d$. Let ϱ_D be a σ-finite measure on D. The kernel $K(\cdot,\cdot)$ is supposed to satisfy the finite trace property*

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

Let $\mathrm{Id} : H(K) \to L_2(D, \varrho_D)$ be the corresponding embedding operator and $(\sigma_k)_{k\in\mathbb{N}}$ the non-increasing sequence of singular numbers. Then we have for $n \in \mathbb{N}$ and

$$m := \left\lfloor \frac{n}{96(\sqrt{2}\log(2n)+5)} \right\rfloor$$

the general bound

$$g_n(\mathrm{Id}) \le 8 \max\left\{ \sigma_m, \sqrt{\frac{\log(8n)}{n} \sum_{j=m}^{\infty} \sigma_j^2} \right\}. \tag{1.2}$$

Note that the maximum in (1.2) is really required since the second quantity may behave better in order than the first one, which, in fact, represents a natural benchmark in our setting. The result refines the bound in [32] as we give precise constants here for the sampling numbers in a very general situation.

What concerns tractability issues, assume for the moment the finiteness of $\sup_d \sum_{j=1}^{\infty} \sigma_{j,d}^{\tau}$ for some $0 < \tau \le 2$ with respect to a d-indexed family of approximation problems, see [41]. As a direct consequence of (1.2) we then obtain *polynomial tractability* with respect to standard information. In addition, the tractability exponent with respect to standard information is arbitrarily close to the exponent with respect to linear information in the framework of Open Problem 127 in [44], see Section 5.4 below. The quantity $T(m)$ from Theorem 1.1 can be replaced in terms of singular values $(\sigma_k)_{k\in\mathbb{N}}$ since we sampled with respect to an "envelope density" adapted to the specific problem and different from ϱ_D. Although we stated the above result only in terms of sampling numbers, the underlying recovery operator is constructive and the determined error bounds hold with high probability. Computing this envelope density (and sample from it) has been studied in [14, Sect. 5]. However, as we will see below, in case of a compactly supported wavelet system or the classical complex Fourier system it is enough to sample from the target density ϱ_D to obtain similar bounds.

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 $\mathbf{L}_m := \left(\sigma_k^{-1}\, e_k(\mathbf{x}^j)\right)_{j=1;\,k=1}^{n;\,m-1} \in \mathbb{C}^{n\times(m-1)}$. In order to solve this linear system of equations, one can apply a standard conjugate gradient type iterative algorithm, e.g., LSQR [47]. 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. Our proof of Theorem 1.1 includes estimates of the singular values of the matrix $\mathbf{L}_m$, which yields that the condition number of $\mathbf{L}_m$ is bounded from above by $\sqrt{3}$ with probability at least $1-\delta$, i.e., only a small number R

of iterations is required. The arithmetic cost of the algorithm will be further reduced if the underlying system matrix $\mathbf{L}_m$ is sparse, which is the case in the wavelet setting when using compactly supported wavelets instead of classical Fourier basis polynomials, see Remark 9.2 below.

As a special case we consider the recovery of functions from periodic Sobolev spaces with mixed smoothness. This problem has been investigated by many authors in the last 30 years, see [16, Sect. 5] and the references therein. There is recent progress in [32], which motivated the investigations in this paper. The authors showed the existence of a point set and a recovery algorithm with improved asymptotic worst-case errors, see Section 8.3 for a detailed discussion of this result. In this paper, we use the simple least squares algorithm from [4, 32], and we show that using random points makes it also possible to obtain explicit worst-case recovery guarantees even in the preasymptotic range. We would like to point out that the general asymptotic as well as the preasymptotic behaviour of the presented recovery guarantees can already be derived from [32]. The contribution of this paper are the explicitly determined, reasonable constants here. The analysis benefits from the fact that the underlying eigenvector system is a bounded orthonormal system (BOS). However, we demonstrate in Section 9 that this assumption is not necessary for our analysis. We show that *Hyperbolic wavelet regression* recovers non-periodic functions belonging to a Sobolev space with mixed smoothness H^s_{mix} from n nodes $\mathbf{X}$ drawn according to ϱ_D in the worst-case with the same asymptotic rate. The proposed approach achieves rates that are only worse by $\log^s n$ in comparison to the optimal rates achieved by hyperbolic wavelet best approximation studied in [19] and [54]. The above general bound on g_n can for instance be used for any non-periodic embedding

$$\mathrm{Id} : H^s_{\text{mix}}([0,1]^d) \to L_2([0,1]^d)\,, \quad s > 1/2\,,$$

where $H^s_{\text{mix}}([0,1]^d)$ 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, 7.4]. Plugging in well-known upper bounds on the singular numbers we improve on the asymptotic sampling bounds by Dinh Dũng [15] and Byrenheid [7]. In addition, using refined preasymptotic estimates for the $(\sigma_j)_{j\in\mathbb{N}}$ in the non-periodic situation (see [30, 4.3]) yields reasonable bounds for g_n in case of small n.

In addition to the worst-case error, we also consider the error for recovering individual functions belonging to reproducing kernel Hilbert spaces. Here we deal with a least squares Monte-Carlo method, which has been already considered in the literature, see for instance [12], [4], [9] and others. We complement the results from [12] and put them into a multivariate context.

A further application of our least squares method is in the field of numerical integration. Oettershagen [45], and also the authors of the recent paper [2], used a least squares optimization approach in order to compute the near-optimal (complex-valued) weights $\mathbf{q} := (q_1, ..., q_n)^\top$ in a cubature formula, i.e.,

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

where $\mathbf{f} := (f(\mathbf{x}^j))_{j=1}^n$ and μ_D is the measure for which we want to compute the integral. This relation between least squares approximation and corresponding cubature rules has been pointed out also by Gröchenig in the recent paper [23]. In our setting, the integration nodes $\mathbf{X} = (\mathbf{x}^1, ..., \mathbf{x}^n)$ are determined once in advance for the whole class. Clearly, the bounds from Theorem 1.1, 1.2 can be literally transferred to estimate the worst-case integration error, see Corollary 8.11 below. In fact, by our refined analysis we were able to settle a conjecture in [45] that the worst-case integration error for $H^s_{\rm mix}(\mathbb{T}^d)$ 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 [45]) 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 by using plain random points. Note that this simple method can compete with most of the quasi-Monte Carlo methods based on lattices and digital nets studied in the literature, see [20, pp. 195, 247]. Moreover, if $s < (d-1)/2$ we get a better asymptotic rate than sparse grid integration which is shown to be in the order of $n^{-s}\log(n)^{(d-1)(s+1/2)}$, see [17].

We practically verify our theoretical findings with several numerical experiments. There we compare the recovery error for the least squares regression method $S^m_{\mathbf{X}}$ compared 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.

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. 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 $\|\mathbf{L}\|_{2\to 2}$. 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 $T: X\to Y$ is a continuous operator we write $\|T\colon X\to Y\|$ 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 $b_n\lesssim a_n$.

2 The setting

We will work in the framework of reproducing kernel Hilbert spaces. The relevant theoretical background can be found in [3, Chapt. 1] and [10, 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

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

for all $\mathbf{x} \in D$. It ensures that point evaluations are continuous functionals on $H(K)$. We will use the notation from [10, Chapt. 4]. In the framework of this paper, the finite trace of the kernel

$$\|K\|_2 := \int_D K(\mathbf{x}, \mathbf{x}) \, \varrho_D(\mathrm{d}\mathbf{x}) < \infty \tag{2.1}$$

or its boundedness

$$\|K\|_\infty := \sup_{\mathbf{x} \in D} \sqrt{K(\mathbf{x}, \mathbf{x})} < \infty \tag{2.2}$$

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

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

With $\ell_\infty(D)$ we denote the set of bounded functions on D and with $\|\cdot\|_{\ell_\infty(D)}$ the supremum norm. Note that we do not need the measure ϱ_D for this embedding.

The embedding operator

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

is compact under the integrability condition (2.1), 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 $\mathrm{Id}^* \circ \mathrm{Id}$ provides an at most countable system of strictly positive eigenvalues $(\lambda_j)_{j \in \mathbb{N}}$. These eigenvalues are summable as a consequence of (2.1) and (2.4), (2.5) 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

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

Then $W_{\varrho_D} := \mathrm{Id}^* \circ \mathrm{Id} : H(K) \to H(K)$ 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.,

$$\lambda_1 \ge \lambda_2 \ge \lambda_3 \ge \cdots > 0 \,.$$

Note that we have by Bessel's inequality

$$\|f\|_{H(K)}^2 \ge \sum_{k=1}^{\infty} |\langle f, e_k \rangle_{H(K)}|^2 \,. \tag{2.3}$$

Let us point out, that by (2.3), the function

$$\mathbf{x} \mapsto \sum_{k=1}^{\infty} |e_k(\mathbf{x})|^2 \le K(\mathbf{x}, \mathbf{x}) \tag{2.4}$$

exists pointwise in $\mathbb{C}$. This implies

$$\sum_{k=1}^{\infty} \lambda_k \le \int_D K(\mathbf{x}, \mathbf{x}) \, \varrho_D(\mathrm{d}\mathbf{x}) \tag{2.5}$$

and we get by $\lambda_j = \sigma_j^2$ that $\mathrm{Id} : H(K) \to L_2(D, \varrho_D)$ is a Hilbert-Schmidt operator if (2.1) holds. We will restrict to the situation where we have equality in (2.5). This can be achieved by posing additional assumptions, namely that $H(K)$ is separable and ϱ_D is σ-finite, see [10, Thm. 4.27]. It further holds that

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

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 two crucial quantities

$$N(m) := \sup_{\mathbf{x}\in D} \sum_{k=1}^{m-1} |\eta_k(\mathbf{x})|^2 \tag{2.6}$$

and

$$T(m) := \sup_{\mathbf{x}\in D} \sum_{k=m}^{\infty} |e_k(\mathbf{x})|^2 \tag{2.7}$$

will play a fundamental role. The first one is often called "spectral function", see [23] and the references therein. Clearly, by (2.5) $N(m)$ and $T(m)$ are well-defined if the kernel is bounded, i.e., if (2.2) 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 $k \in \mathbb{N}$

$$\|\eta_k\|_{\ell_\infty(D)} \le B \,. \tag{2.8}$$

Let us call B the BOS constant of the system. In this case we have

$$N(m) \le (m-1)B^2 \quad \text{and} \quad T(m) \le B^2 \sum_{k=m}^{\infty} \lambda_k \,.$$

Remark 2.1. We would like to point out an issue concerning the embedding operator $\mathrm{Id} : H(K) \to L_2(D, \varrho_D)$ defined above. As discussed in [10, 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 [10, 4.5] and [10, Ex. 4.6, p. 163], which is related to Mercer's theorem [10, 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

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

as well as equality signs in (2.3), (2.4) and (2.5), 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 [10, 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 5.7) holds already true under the finite trace condition (2.1) 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 [26] 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 $\tau_1, ..., \tau_m > 0$ arranged in non-increasing order.*

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

(ii) *In particular, it holds that*

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

whenever $\mathbf{L} = \mathbf{U}^*\mathbf{\Sigma}\mathbf{V}$, where $\mathbf{\Sigma} \in \mathbb{R}^{n\times m}$ is a rectangular matrix only with $(\tau_1, ..., \tau_m)$ 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 $\tilde{\mathbf{\Sigma}} \in \mathbb{R}^{m\times n}$ denotes the matrix with $(\tau_1^{-1}, ..., \tau_m^{-1})$ on the "main diagonal".

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

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

Proof. We start with $\mathbf{L} = \mathbf{U}^*\mathbf{\Sigma}\mathbf{V}$ and write

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

We finally put

$$\tilde{\mathbf{\Sigma}} := (\mathbf{\Sigma}^*\mathbf{\Sigma})^{-1}\mathbf{\Sigma}^*$$

and immediately observe the desired properties. □

For function recovery we will use the following matrix

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

for $\mathbf{X} = (\mathbf{x}^1, ..., \mathbf{x}^n) \in D^n$ 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 $\mathbf{x}^1, ..., \mathbf{x}^n$ are chosen independently and identically ϱ_D-distributed from D.

Algorithm 1 Least squares regression.

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

Solve the over-determined linear system

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

via least squares (e.g. directly or via the LSQR algorithm [47]), i.e., compute

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

Output: $\mathbf{c} = (c_1, ..., c_{m-1})^\top \in \mathbb{C}^{m-1}$ 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, $k = 1, \ldots, m-1$, of the approximant

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

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

$$\mathbf{X} = (\mathbf{x}^1, ..., \mathbf{x}^n) \in D^n.$$

4 Concentration inequalities

We will consider complex-valued random variables X and random vectors $(X_1, ..., X_N)$ 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

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

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,

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

Let us state concentration inequalities for the norm of sums of complex rank-1 matrices. For the first result we refer to Oliveira [46]. 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

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

Proposition 4.1. *Let $\mathbf{y}^i, i = 1, ..., n$, 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 $\mathbb{E}(\mathbf{y}^i \otimes \mathbf{y}^i) = \mathbf{\Lambda} \in \mathbb{C}^{N\times N}$ and $0 < t < 1$. Then it holds*

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

Proof. In order to show the probability estimate, we refer to the proof of [46, Lem. 1] and observe

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

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. □

Remark 4.2. A slightly stronger version for the case of real matrices can be found in Cohen, Davenport, Leviatan [12] (see also the correction in [13]). For $\mathbf{y}^i, i = 1, ..., n$, 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 [37], and Rudelson [52]. The complex version with precise constants can be found in Rauhut [49, Cor. 6.20].

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

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

with

$$C_{\mathrm{R}} = \sqrt{2} + \frac{1}{4\sqrt{2\log(8)}} \in [1.53, 1.54]\,. \tag{4.4}$$

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 [49, Lem. 6.18] which itself is based on a noncommutative Khintchine inequality, see [49, 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 [49, Thm. 6.14] are replaced by rank-1-operators $\mathbf{B}_j\colon \ell_2 \to \ell_2$ 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 noncommutative Khintchine inequality is available. This implies an ℓ_2-version of [49, Lem. 6.18] which reads as follows: Let $\mathbf{y}^j \in \ell_2$, $j = 1, ..., n$, and $p \geq 2$. Then

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

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 ([49, 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 $\mathbf{y}^i$, $i = 1, ..., n$ 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*

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

Proof. By well-known symmetrization technique (see [21, Lem. 8.4]), Proposition 4.3, and the Cauchy-Schwarz inequality, we obtain

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

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

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

and this corresponds to the assertion. □

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 $\mathbf{x}^1, ..., \mathbf{x}^n \in D$ 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, 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 $\{\exp(2\pi \mathrm{i}\, \mathbf{k}\cdot\mathbf{x}) \colon \mathbf{k} \in \mathbb{Z}^d\}$ 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 $\{\eta_1(\cdot), \eta_2(\cdot), \eta_3(\cdot), ...\}$ be the orthonormal system in $L_2(D, \varrho_D)$ induced by the kernel K and $\mathbf{x}^1, ..., \mathbf{x}^n \in D$ 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 (2.6) *and* $\mathbf{I}_m = \mathrm{diag}(\mathbf{1}) \in \{0,1\}^{(m-1)\times(m-1)}$.

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

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

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

So, in case of a bounded orthonormal system with BOS constant $B > 0$, we may choose

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

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 $\tau_1, ..., \tau_{m-1}$ 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 $\tau_1, ..., \tau_{m-1} \geq \sqrt{n/2}$. 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 $\{\eta_1(\cdot), \eta_2(\cdot), \eta_3(\cdot), ...\}$ 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 [32], 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 $\mathbf{x}^1, ..., \mathbf{x}^n$ be drawn uniformly and i.i.d. at random according to ϱ_D, and consider the n i.i.d. random sequences*

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

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

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

has expected norm

$$\mathbb{E}(\|\mathbf{\Phi}_m\|^2) \leq n \left(\sigma_m^2 + 4\, C_{\mathrm{R}}^2 \frac{\log(8n)}{n} T(m) + 2\, C_{\mathrm{R}}\, \sigma_m \sqrt{\frac{\log(8n)}{n} T(m)} \right). \tag{5.3}$$

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

$$\mathbf{\Lambda}_m := \mathbb{E}\Big(\frac{1}{n} \mathbf{\Phi}_m^* \mathbf{\Phi}_m\Big) = \mathbb{E}(\mathbf{y}^i \otimes \mathbf{y}^i) = \mathrm{diag}(\sigma_m^2, \sigma_{m+1}^2, ...).$$

This gives

$$\|\mathbf{\Phi}_m\|^2 = \|\mathbf{\Phi}_m^* \mathbf{\Phi}_m\| \leq \|\mathbf{\Phi}_m^* \mathbf{\Phi}_m - n\mathbf{\Lambda}_m\| + n\|\mathbf{\Lambda}_m\|.$$

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

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 $\mathrm{Id} : H(K) \to L_2(D, \varrho_D)$ 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 $(\mathbf{x}^1, ..., \mathbf{x}^n)$ i.i.d. at random according to ϱ_D, we have for the conditional expectation of the worst-case error*

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

with C_{R} from (4.4).

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

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

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 $\mathbf{x} \in D$

$$
\begin{aligned}
\Big(f - P_{m-1} f \Big)(\mathbf{x}) &= \langle f, K(\cdot, \mathbf{x}) - \sum_{j=1}^\infty e_j(\cdot) \overline{e_j(\mathbf{x})} + \sum_{j=1}^\infty e_j(\cdot) \overline{e_j(\mathbf{x})} - \sum_{j=1}^{m-1} e_j(\cdot) \overline{e_j(\mathbf{x})} \rangle_{H(K)} \\
&= \sum_{j=m}^\infty \langle f, e_j \rangle_{H(K)} e_j(\mathbf{x}) + \langle f, T(\cdot, \mathbf{x}) \rangle_{H(K)} ,
\end{aligned} \tag{5.6}
$$

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

$$
\|T(\cdot, \mathbf{x})\|^2_{H(K)} := \langle T(\cdot, \mathbf{x}), T(\cdot, \mathbf{x}) \rangle_{H(K)} = K(\mathbf{x}, \mathbf{x}) - \sum_{j=1}^\infty |e_j(\mathbf{x})|^2 . \tag{5.7}
$$

This gives

$$
\begin{aligned}
\Big| \Big(f - P_{m-1} f \Big)(\mathbf{x}) \Big|^2 &\leq \Big| \sum_{j=m}^\infty \langle f, e_j \rangle_{H(K)} e_j(\mathbf{x}) \Big|^2 + 2 \|f\|^2_{H(K)} \|T(\cdot, \mathbf{x})\|_{H(K)} \sqrt{\sum_{j=m}^\infty |e_j(\mathbf{x})|^2} \\
&\quad + \|f\|^2_{H(K)} \|T(\cdot, \mathbf{x})\|^2_{H(K)} \\
&\overset{(2.4)}{\leq} \left(\sum_{i=m}^\infty \overline{\langle f, e_i \rangle_{H(K)} e_i(\mathbf{x})} \right) \left(\sum_{j=m}^\infty \langle f, e_j \rangle_{H(K)} e_j(\mathbf{x}) \right) \\
&\quad + \|f\|^2_{H(K)} \|T(\cdot, \mathbf{x})\|_{H(K)} \Big(\|T(\cdot, \mathbf{x})\|_{H(K)} + 2 \|K\|_\infty \Big) \\
&\leq \sum_{i=m}^\infty \sum_{j=m}^\infty \overline{\langle f, e_i \rangle_{H(K)}} \langle f, e_j \rangle_{H(K)} \overline{e_i(\mathbf{x})} e_j(\mathbf{x}) \\
&\quad + 3 \|T(\cdot, \mathbf{x})\|_{H(K)} \|K\|_\infty \|f\|^2_{H(K)} .
\end{aligned}
$$

Returning to (5.5) we estimate

$$
\begin{aligned}
\sum_{k=1}^n \Big| \Big(f - P_{m-1} f \Big)(\mathbf{x}^k) \Big|^2 &\leq \|(\langle f, e_j \rangle_{H(K)})_{j \in \mathbb{N}}\|_2^2 \, \|\mathbf{\Phi}_m\|^2 + 3 \|K\|_\infty \|f\|^2_{H(K)} \sum_{k=1}^n \|T(\cdot, \mathbf{x}^k)\| \\
&\leq \|f\|^2_{H(K)} \left(\|\mathbf{\Phi}_m\|^2 + 3 \|K\|_\infty \sum_{k=1}^n \|T(\cdot, \mathbf{x}^k)\| \right) ,
\end{aligned} \tag{5.8}
$$

where $\mathbf{\Phi}_m$ denotes the infinite matrix from Proposition 5.4. Note that we used (2.3) in the last but one step. The relation in (5.8) together with (5.5) and $\|f\|_{H(K)} \le 1$ implies

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

Integrating on both sides yields

$$
\begin{aligned}
&\int\limits_{\|\mathbf{H}_m - \mathbf{I}_m\| \le 1/2} \sup_{\|f\|_{H(K)} \le 1} \|f - S^m_{\mathbf{X}} f\|^2_{L_2(D,\varrho_D)} \, \varrho_D^n(\mathrm{d}\mathbf{X}) \\
&\quad \le \sigma_m^2 + \frac{2}{n} \mathbb{E}(\|\mathbf{\Phi}_m\|^2) + \frac{6\|K\|_\infty}{n} \sum_{k=1}^{n} \int_D \|T(\cdot, \mathbf{x}^k)\| \, \varrho_D(\mathrm{d}\mathbf{x}) \qquad (5.9) \\
&\quad = \sigma_m^2 + \frac{2}{n} \mathbb{E}(\|\mathbf{\Phi}_m\|^2) \, .
\end{aligned}
$$

Note that the integral on the right-hand side of (5.9) vanishes because of (5.7) and the fact that we have an equality sign in (2.5) due to our assumptions (separability of $H(K)$). This gives

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

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

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 $\delta > 0$*

$$
\mathbb{P}\left(\sup_{\|f\|_{H(K)} \le 1} \|f - S^m_{\mathbf{X}} f\|^2_{L_2(D,\varrho_D)} \le \frac{C}{\delta} \max\left\{ \sigma_m^2, \frac{\log(8n)}{n} T(m) \right\} \right) \ge 1 - 2\delta \, , \qquad (5.10)
$$

where $C := 3 + 8\, C_{\mathrm{R}}^2 + 4\, C_{\mathrm{R}} < 28.05$ is an absolute constant.

Proof. We define the events

$$
\begin{aligned}
A &:= \left\{ \mathbf{X} : \sup_{\|f\|_{H(K)} \le 1} \|f - S^m_{\mathbf{X}} f\|^2_{L_2(D,\varrho_D)} \le t \right\} , \\
B &:= \{ \mathbf{X} : \|\mathbf{H}_m - \mathbf{I}_m\| \le 1/2 \}
\end{aligned}
$$

and split up

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

Treating each summand separately, we have

$$
\begin{aligned}
\mathbb{P}(A^{\complement} \cap B) &= \mathbb{P}(A^{\complement} | B) \mathbb{P}(B) \le \mathbb{P}(A^{\complement} | B)(1 - \delta) \, , \\
\mathbb{P}(A^{\complement} \cap B^{\complement}) &\le \mathbb{P}(B^{\complement}) \le \delta \, .
\end{aligned}
$$

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

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

which yields

$$\mathbb{P}(A^{\complement}|B) \le \frac{\mathbb{E}(A^{\complement}|B)}{t} \le \frac{\delta}{1-\delta}$$

and the assertion follows. □

5.3 Optimal sampling recovery of functions

We are interested in the question of optimal sampling recovery of functions from reproducing kernel Hilbert spaces in $L_2(D, \varrho_D)$. The quantity we want to study is classically given by

$$g_n(\mathrm{Id} : H(K) \to L_2(D, \varrho_D)) := \inf_{\mathbf{X} = (\mathbf{x}^1, \ldots, \mathbf{x}^n)} \inf_{\substack{R : \mathbb{C}^n \to L_2 \\ \text{linear}}} \sup_{\|f\|_{H(K)} \le 1} \|f - R(f(\mathbf{X}))\|_{L_2(D, \varrho_D)}\,.$$

The goal is to get reasonable bounds for this quantity in n, preferably in terms of the singular numbers of the embedding. Theorem 5.5 already gives an answer if the kernel is bounded and the eigenfunctions $(\sigma_k^{-1} e_k(\cdot))_{k\in\mathbb{N}}$ are uniformly bounded in $\ell_\infty(D)$. In order to drop both conditions, we will use a reweighted (deterministic) least squares algorithm (Algorithm 2 below) to recover functions $f \in H(K)$ from samples at random nodes ("random information" in the sense of [25]). This algorithm is a modification of the one proposed earlier in [32]. Note that both algorithms coincide in case of an injective embedding $\mathrm{Id} : H(K) \to L_2(D, \varrho_D)$, see Remark 2.1 above. 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 [14], see Remark 6.3 below. Also in connection with compressed sensing it led to substantial improvements when recovering multivariate functions, see [51, 50]. Authors originally applied this technique e.g. for the approximation of integrals, see [24]. However, the setting in which we are interested in requires additional work since the sampling nodes are drawn for the whole class simultaneously.

Theorem 5.7. *Let $H(K)$ be a separable reproducing kernel Hilbert space of complex-valued functions defined on D such that*

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

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 $\mathrm{Id} : H(K) \to L_2(D, \varrho_D)$*. Then we have for $n \in \mathbb{N}$ and*

$$m := \left\lfloor \frac{n}{96(\sqrt{2}\log(2n) + 5)} \right\rfloor \tag{5.11}$$

the general bound

$$g_n(\mathrm{Id})^2 \le 50 \max\left\{\sigma_m^2, \frac{\log(8n)}{n} \sum_{j=m}^{\infty} \sigma_j^2\right\}\,. \tag{5.12}$$

Proof. We emphasize that the right hand side in (5.12) makes sense since we know from (2.5) that the sequence of singular numbers is square-summable. As a modification of the density function which has been presented in [32] we use the density $\varrho_m(\mathbf{x}) : D \to \mathbb{R}$, defined for any $m \in \mathbb{N}$ as

$$\varrho_m(\mathbf{x}) := \frac{1}{2}\left(\frac{1}{m-1}\sum_{j=1}^{m-1}|\eta_j(\mathbf{x})|^2 + \left(\sum_{j=m}^{\infty}\lambda_j\right)^{-1}\left(K(\mathbf{x},\mathbf{x}) - \sum_{j=1}^{m-1}|e_j(\mathbf{x})|^2\right)\right). \tag{5.13}$$

As above, the family $(e_j(\cdot))_{j\in\mathbb{N}}$ represents the eigenvectors of the non-vanishing eigenvalues of the compact self-adjoint operator $W_{\varrho_D} := \mathrm{Id}^* \circ \mathrm{Id} : H(K) \to H(K)$, the sequence $(\lambda_j)_{j\in\mathbb{N}}$ represents the ordered eigenvalues, and finally $\eta_j := \lambda_j^{-1/2} e_j$.

Clearly, as a consequence of (2.4) 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. Since we assume the separability of $H(K)$ and the σ-finiteness of ϱ_D we observe equality in (2.5), cf. [10, Thm. 4.27], and thus we easily see that $\int_D \varrho_m(\mathbf{x})\,\varrho_D(\mathrm{d}\mathbf{x}) = 1$. Let us define a family of kernels $\widetilde{K}_m(\mathbf{x},\mathbf{y})$, indexed by $m \in \mathbb{N}$, via

$$\widetilde{K}_m(\mathbf{x},\mathbf{y}) := \frac{K(\mathbf{x},\mathbf{y})}{\sqrt{\varrho_m(\mathbf{x})\varrho_m(\mathbf{y})}}\,, \tag{5.14}$$

and a new measure

$$\mu_m(A) = \int_A \varrho_m(\mathbf{x})\,\varrho_D(\mathrm{d}\mathbf{x})\,, \quad A \in \mathcal{A}\,, \tag{5.15}$$

with the corresponding weighted space $L_2(D,\mu_m)$. Clearly, $\widetilde{K}_m(\cdot,\cdot)$ is a positive type function. As a consequence of

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

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

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

Indeed,

$$\begin{aligned} K(\mathbf{x},\mathbf{x}) &= \sum_{j=1}^{m-1}|e_j(\mathbf{x})|^2 + \left(K(\mathbf{x},\mathbf{x}) - \sum_{j=1}^{m-1}|e_j(\mathbf{x})|^2\right) \\ &= \sum_{j=1}^{m-1}\lambda_j|\eta_j(\mathbf{x})|^2 + \left(\sum_{j=m}^{\infty}\lambda_j\right)\cdot\left(\sum_{j=m}^{\infty}\lambda_j\right)^{-1}\left(K(\mathbf{x},\mathbf{x}) - \sum_{j=1}^{m-1}|e_j(\mathbf{x})|^2\right) \\ &\le c_m\varrho_m(\mathbf{x}) \end{aligned} \tag{5.17}$$

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 $\widetilde{K}_m(\mathbf{x},\mathbf{y}) := 0$ in (5.14). In any case, due to (5.16), the kernel $\widetilde{K}_m(\mathbf{x},\mathbf{y})$ fits the requirements in Theorem 5.5. In fact, in Theorem 5.5 it is necessary that $\widetilde{N}(m)$ and $\widetilde{T}(m)$ are well-defined and that we have access to function values in order to create the matrices $\widetilde{\mathbf{L}}_m$ and take the function values $f(\mathbf{x}^k)$,

$k = 1, ..., n$. Let us discuss the quantities $\widetilde{N}(m)$ and $\widetilde{T}(m)$ appearing in Theorem 5.5 for this new kernel $\widetilde{K}_m(\mathbf{x}, \mathbf{y})$ first. It is clear, that the embedding $\mathrm{Id} : H(\widetilde{K}_m) \to L_2(D, \mu_m)$ shares the same spectral properties as the original embedding. Note that a function g belongs to $H(\widetilde{K}_m)$ 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 (5.18) 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 $\|g\|_{H(\widetilde{K}_m)} \leq 1$. Indeed, let

$$f(\cdot) = \sum_{i=1}^{N} \alpha_i K(\cdot, \mathbf{x}^i) \tag{5.18}$$

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

$$\begin{aligned} g = f(\cdot)/\sqrt{\varrho_m(\cdot)} &= \sum_{i=1}^{N} \alpha_i K(\cdot, \mathbf{x}^i)/\sqrt{\varrho_m(\cdot)} \\ &= \sum_{i=1}^{N} \alpha_i \sqrt{\varrho_m(\mathbf{x}^i)} \frac{K(\cdot, \mathbf{x}^i)}{\sqrt{\varrho_m(\cdot)}\sqrt{\varrho_m(\mathbf{x}^i)}} \, . \end{aligned} \tag{5.19}$$

This implies

$$\begin{aligned} \langle g, g \rangle_{H(\widetilde{K}_m)} &= \sum_{j=1}^{N} \sum_{i=1}^{N} \alpha_i \overline{\alpha_j} \sqrt{\varrho_m(\mathbf{x}^i)} \sqrt{\varrho_m(\mathbf{x}^j)} \frac{K(\mathbf{x}^j, \mathbf{x}^i)}{\sqrt{\varrho_m(\mathbf{x}^j)}\sqrt{\varrho_m(\mathbf{x}^i)}} \\ &= \langle f, f \rangle_{H(K)} \, . \end{aligned} \tag{5.20}$$

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,

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

Furthermore, taking (2.4) into account, we find

$$\begin{aligned} \widetilde{T}(m) &:= \sup_{\mathbf{x} \in D} \sum_{k=m}^{\infty} |e_k(\mathbf{x})|^2/\varrho_m(\mathbf{x}) \\ &\leq 2\Bigg(\sum_{j=m}^{\infty} \lambda_j\Bigg) \sup_{\mathbf{x} \in D} \sum_{k=m}^{\infty} |e_k(\mathbf{x})|^2 \Bigg/ \Bigg(K(\mathbf{x}, \mathbf{x}) - \sum_{j=1}^{m-1} |e_j(\mathbf{x})|^2\Bigg) \\ &\leq 2\Bigg(\sum_{j=m}^{\infty} \lambda_j\Bigg) \sup_{\mathbf{x} \in D} \sum_{k=m}^{\infty} |e_k(\mathbf{x})|^2 \Bigg/ \sum_{j=m}^{\infty} |e_j(\mathbf{x})|^2 \\ &\leq 2 \sum_{j=m}^{\infty} \lambda_j = 2 \sum_{j=m}^{\infty} \sigma_j^2 \, . \end{aligned} \tag{5.21}$$

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

$$\widetilde{\mathbf{L}}_m \tilde{\mathbf{c}} = \boldsymbol{g}, \text{ where } \widetilde{\mathbf{L}}_m := \Big(\tilde{\eta}_j(\mathbf{x}^k)\Big)_{k=1,j=1}^{n,m-1}, \ \boldsymbol{g} := \left(\frac{f(\mathbf{x}^1)}{\sqrt{\varrho_m(\mathbf{x}^1)}}, \dots, \frac{f(\mathbf{x}^n)}{\sqrt{\varrho_m(\mathbf{x}^n)}}\right)^\top, \tag{5.22}$$

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

$$\|g - \widetilde{S}^m_{\mathbf{X}}(g)\|_{L_2(D,\mu_m)} = \|f - \sqrt{\varrho_m}\, \widetilde{S}^m_{\mathbf{X}}(g)\|_{L_2(D,\varrho_D)} = \|f - \widetilde{S}^m_{\mathbf{X}} f\|_{L_2(D,\varrho_D)}\,, \tag{5.23}$$

where $\widetilde{S}^m_{\mathbf{X}} f := \sqrt{\varrho_m}\, \widetilde{S}^m_{\mathbf{X}}(g) = \sum_{j=1}^{m-1} \tilde{c}_j \eta_j(\mathbf{x})$.

Algorithm 2 (Reweighted) least squares regression.

Input:	$\mathbf{X} = (\mathbf{x}^1, ..., \mathbf{x}^n) \in D^n$	set of distinct sampling nodes,
	$\mathbf{f} = (f(\mathbf{x}^1), ..., f(\mathbf{x}^n))^\top$	samples of f evaluated at the nodes from $\mathbf{X}$,
	$m \in \mathbb{N}$	$m < n$ such that the matrix $\tilde{\mathbf{L}}_m$ in (5.24) has full (column) rank.

Compute reweighted samples $\boldsymbol{g} := (g_k)_{k=1}^n$ 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

$$\widetilde{\mathbf{L}}_m\, (\tilde{c}_1, ..., \tilde{c}_{m-1})^\top = \mathbf{g}\,,\ \widetilde{\mathbf{L}}_m := \big(l_{k,j}\big)_{k=1,j=1}^{n,m-1},\ l_{k,j} := \begin{cases} 0, & \varrho_m(\mathbf{x}^k) = 0, \\ \eta_j(\mathbf{x}^k)/\sqrt{\varrho_m(\mathbf{x}^k)}, & \varrho_m(\mathbf{x}^k) \neq 0, \end{cases} \tag{5.24}$$

via least squares (e.g. directly or via the LSQR algorithm [47]), i.e., compute

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

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

We stress that $\widetilde{S}^m_{\mathbf{X}} f$ and the direct computation of $S^m_{\mathbf{X}} f$ 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 (5.11) we have for $\mathbf{X} = (\mathbf{x}^1, ..., \mathbf{x}^n)$ the relation

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

Applying Theorem 5.5 with $\delta = \mathrm{e}^{-5}$ yields

$$\begin{aligned}
\mathbb{E}&\left(\sup_{\|f\|_{H(K)}\le 1} \|f - S^m_{\mathbf{X}} f\|^2_{L_2(D,\varrho_D)} \Big| \|n^{-1}\widetilde{\mathbf{L}}^*_m \widetilde{\mathbf{L}}_m - \mathbf{I}_m\| \le 1/2 \right) \\
&\le \frac{1}{1-\mathrm{e}^{-5}} \left(3\sigma_m^2 + 8\, C_{\mathrm{R}}^2 \frac{\log(8n)}{n} \widetilde{T}(m) + 4\, C_{\mathrm{R}}\, \sigma_m \sqrt{\frac{\log(8n)}{n}\widetilde{T}(m)} \right) \\
&\le \frac{1}{1-\mathrm{e}^{-5}} \left(3\sigma_m^2 + 16\, C_{\mathrm{R}}^2 \frac{\log(8n)}{n} \sum_{j=m}^{\infty} \sigma_j^2 + 4\sqrt{2}\, C_{\mathrm{R}}\, \sigma_m \sqrt{\frac{\log(8n)}{n} \sum_{j=m}^{\infty} \sigma_j^2} \right) \\
&\le \frac{3 + 16\, C_{\mathrm{R}}^2 + 4\sqrt{2}\, C_{\mathrm{R}}}{1-\mathrm{e}^{-5}} \max\left\{ \sigma_m^2, \frac{\log(8n)}{n} \sum_{j=m}^{\infty} \sigma_j^2 \right\} \le 50 \cdot \max\left\{ \sigma_m^2, \frac{\log(8n)}{n} \sum_{j=m}^{\infty} \sigma_j^2 \right\}
\end{aligned}$$

if

$$\tilde{N}(m) \le 2m \le \frac{n}{48(\sqrt{2}\log(2n) + 5)}\,,$$

where in the last but one estimate we took (5.21) into account. Finally, the estimate on the conditional expectation implies the existence of a set $\mathbf{X} = (\mathbf{x}^1, ..., \mathbf{x}^n)$ such that the above bound holds. □

Remark 5.8. Algorithm 2 is a modification of the one proposed earlier in [32]. Note that both algorithms coincide in case of an injective embedding $\mathrm{Id} : H(K) \to L_2(D, \varrho_D)$, see Remark 2.1 above. In order to prove the theorem in full generality, we use a modification of the density function presented in [32]. The pointwise defined density function will be an essential ingredient for getting the nodes $\mathbf{X} = (\mathbf{x}^1, ..., \mathbf{x}^n)$ 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 $H(\widetilde{K}_m)$ built upon the modified density function. To this situation Theorem 5.5 is applied, and we obtain Theorem 5.7, which also improves the results in [32] by determining explicit constants.

5.4 The power of standard information

Assuming that we have a polynomial decay of the $(\sigma_n)_n$, i.e., there are positive numbers $\alpha > 1/2, \beta > 0$ such that

$$\sigma_n \lesssim_d n^{-\alpha} (\log n)^{\beta}\,,$$

we obtain a similar bound for the sampling numbers. Indeed, from Theorem 5.7 we obtain by an elementary calculation

$$g_n \lesssim_d n^{-\alpha} (\log n)^{\alpha+\beta}\,.$$

In other words, it follows that $\alpha^{\mathrm{std}}_{2,\varrho_D} = \alpha^{\mathrm{lin}}_{2,\varrho_D}$, i.e., the polynomial order of convergence for standard information is the same as for linear information if the kernel has a finite trace (i.e., $\mathrm{Id} : H(K) \to L_2(D, \varrho_D)$ is a Hilbert-Schmidt operator). This problem has been addressed in [36], [43, Open Problem 1] and [44, Open Problem 126]. In [32] 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 5.7 is to get explicitly determined constants on the one hand. On the other hand, it shows that the finite trace condition is essentially enough for this purpose. Note that a finite trace is natural in this context, see [26].

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 [41, 42, 44]. We comment on polynomial tractability with respect to *linear information* Λ^{all} and *standard information* Λ^{std}. We consider the family of approximation problems

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

where $K_d : D_d \times D_d \to \mathbb{C}$ is a family of reproducing kernels. In [41, 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

$$C := \sup_d \left(\sum_{j=1}^{\infty} \sigma_{j,d}^{\tau} \right)^{1/\tau} < \infty, \tag{5.25}$$

where $\sigma_{j,d}$, $j = 1, \ldots$, are the singular values belonging to APP_d for fixed d. From our analysis in Theorem 5.7 above we directly obtain a sufficient condition for polynomial tractability with respect to Λ^{std}.

Theorem 5.9. *The family* $\{\text{APP}_d\}$ *is strongly polynomially tractable with respect to* Λ^{std}

(i) if there exists $\tau \le 2$ *such that* (5.25) *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.,* $n^{\text{wor}}(\varepsilon, d; \Lambda^{\text{all}}) \le C_{\text{all}}\, \varepsilon^{-p_{\text{all}}}$.

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

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

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 5.7 implies strong polynomial tractability if (i) is assumed. In case (i) we may use Stechkin's lemma [16, Lem. 7.8] which gives that $\sum_{j=m}^{\infty} \sigma_{j,d}^2 \le C^2 m^{-2/\tau+1}$ for all d. This gives the exponent $p_{\text{std}} = \tau$ and an additional log due to (5.11). If (ii) is assumed then $\sum_{j=m}^{\infty} \sigma_{j,d}^2 \le C' m^{-2/p_{\text{all}}+1}$ for all d. □

Theorem 5.9 is stronger than [44, Thm. 26.20] in two aspects. As pointed out in the proof, assumption (i) is weaker than (ii), which is essentially the one in [44, Thm. 26.20]. Furthermore, our statement is stronger since p_{std} equals p_{all}. The authors in [44, 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 5.10. Let us consider an example from [48] and [35], namely, if $s_1 \le s_2 \le ... \le s_d$ then

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

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

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

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

$$s_j \geq \beta \log_2(j+1)\,, \quad j \in \mathbb{N}\,. \tag{5.26}$$

It has been shown in [48], see also [35], that the growth condition (5.26) is necessary and sufficient for polynomial tractability with respect to Λ^{all}. Taking Theorem 5.9 into account we easily check that (5.25) 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

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

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 [41, 42, 44], such a scenario is called "randomized setting", which refers to the occurrence of a random element in the algorithm (Monte Carlo). In [59], the authors presented a constructive, randomized algorithm based on random samples for which one expects approximation errors converging with a rate arbitrarily close to the rate by which $(\sigma_k)_{k\in\mathbb{N}}$ converges to zero. To this end, the authors construct a number of different probability density functions which are used in order to randomly choose the sampling nodes. Furthermore, they discuss simplifications of this algorithm that provide the same order of convergence or an order of convergence of $1/2$ in cases where $(\eta_k)_{k\in\mathbb{N}}$ is a BOS, cf. (5.16), or where the kernel K is bounded, cf. (2.2), respectively. The following investigations improve these results from two 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 is 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 [12, 13]. With similar techniques as above we will get an estimate for the conditional mean of the individual error $\|f - S^m_{\mathbf{X}} f\|_{L_2(D,\varrho_D)}$.

Theorem 6.1. *Let ϱ_D be a probability measure on D and $H(K)$ denote a reproducing kernel Hilbert space of functions on D such that*

$$\sup_{\mathbf{x}\in D} K(\mathbf{x},\mathbf{x}) < \infty\,. \tag{6.1}$$

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 $\{\mathbf{x}^1, ..., \mathbf{x}^n\} \subset D$ i.i.d. at random according to ϱ_D, we have for the conditional expectation of the individual error

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

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

Proof. This time we follow the proof of [12, Thm. 2] and obtain

$$
\begin{aligned}
\|f - S_{\mathbf{X}}^m f\|_{L_2(D,\varrho_D)}^2 &= \|f - P_{m-1} f\|_{L_2(D,\varrho_D)}^2 + \|P_{m-1} f - S_{\mathbf{X}}^m f\|_{L_2(D,\varrho_D)}^2 \\
&\le \sigma_m^2 + \|S_{\mathbf{X}}^m (P_{m-1} f - f)\|_{L_2(D,\varrho_D)}^2 \\
&\le \sigma_m^2 + \|(\mathbf{L}_m^* \mathbf{L}_m)^{-1}\|^2 \, \|\mathbf{L}_m^* \, (g(\mathbf{x}^1), ..., g(\mathbf{x}^n))^\top\|_2^2 \\
&\le \sigma_m^2 + \frac{4}{n^2} \sum_{k=1}^{m-1} \left| \sum_{j=1}^{n} \overline{\eta_k(\mathbf{x}^j)} g(\mathbf{x}^j) \right|^2 \\
&= \sigma_m^2 + \frac{4}{n^2} \sum_{k=1}^{m-1} \sum_{j=1}^{n} \sum_{i=1}^{n} g(\mathbf{x}^i) \overline{g(\mathbf{x}^j) \eta_k(\mathbf{x}^i)} \eta_k(\mathbf{x}^j) \, ,
\end{aligned}
\tag{6.2}
$$

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

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

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

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

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

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 $\delta > 0$*

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

Remark 6.3. Following [14] we may relax requirement (6.1) on the kernel K to

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

(where ϱ_D is also not longer a probability measure) and improve condition (2.6) to $\widetilde{N}(m) = m - 1$, i.e., we can choose

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

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

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

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 least squares problem similar to that we discussed in Section 5.3.

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 $\mathrm{Q}_{\mathbf{X}}^m$

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

where the weights q_j are determined by $\mathbf{X} = (\mathbf{x}^1, ..., \mathbf{x}^n)$. Indeed, assuming $\mathbf{L}_m$ of full column rank and following Oettershagen [45], we set

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

In fact, the cubature rule $\mathrm{Q}_{\mathbf{X}}^m$ 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), $\boldsymbol{c} := (\mathbf{L}_m^*\,\mathbf{L}_m)^{-1}\,\mathbf{L}_m^*\,\boldsymbol{f}$, since

$$\begin{aligned} \mathrm{Q}_{\mathbf{X}}^m f &= \sum_{j=1}^n q_j f(\mathbf{x}^j) = \boldsymbol{q}^\top \boldsymbol{f} = (\overline{\mathbf{L}_m}\,\overline{(\mathbf{L}_m^*\,\mathbf{L}_m)^{-1}\,\boldsymbol{b}})^\top \boldsymbol{f} = \boldsymbol{b}^\top (\mathbf{L}_m^*\,\mathbf{L}_m)^{-1}\,\mathbf{L}_m^*\,\boldsymbol{f}\\ &= \sum_{k=1}^{m-1} c_k \int_D \eta_k\,\mathrm{d}\mu_D = \int_D S_{\mathbf{X}}^m f\,\mathrm{d}\mu_D\,. \end{aligned}$$

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

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*

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

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

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

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

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

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 $\int_D K(\mathbf{x},\mathbf{x})\,\varrho_D(\mathrm{d}\mathbf{x}) < \infty$ and that ϱ_D is a finite measure. We define the following modified (reweighted) cubature formula

$$\begin{aligned}\widetilde{Q}^m_{\mathbf{X}} f &:= \int_D \widetilde{S}^m_{\mathbf{X}} f(\mathbf{x}) \varrho_D(\mathbf{x}) = \widetilde{\boldsymbol{q}}^\top \boldsymbol{f} \\ &= (\mathbf{D}_{\varrho_m} \overline{\widetilde{\mathbf{L}}_m}\, \overline{(\widetilde{\mathbf{L}}^*_m \widetilde{\mathbf{L}}_m)^{-1} \boldsymbol{b}})^\top \boldsymbol{f} = \boldsymbol{b}^\top (\widetilde{\mathbf{L}}^*_m \widetilde{\mathbf{L}}_m)^{-1} \widetilde{\mathbf{L}}^*_m \mathbf{D}_{\varrho_m} \boldsymbol{f}\,,\end{aligned} \tag{7.3}$$

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

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

Here $\varrho_m(\cdot)$ is given by (5.13) above and depends on the spectral properties of the kernel. Note that we simply put the respective entry to 0 in $\mathbf{D}_{\varrho_m}$ if $\varrho_m(\mathbf{x}^n)$ happens to be zero.

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

$$\sup_{\|f\|_{H(K)}\leq 1} \left| \int_D f(\mathbf{x})\,\varrho_D(\mathrm{d}\mathbf{x}) - \widetilde{Q}^m_{\mathbf{X}}(f)\right|^2 \leq 50 \varrho_D(D) \max\left\{\sigma^2_m, \frac{\log(8n)}{n} \sum_{j=m}^{\infty} \sigma_j^2\right\} \tag{7.4}$$

holds with high probability if the nodes $\mathbf{X} = (\mathbf{x}^1, ..., \mathbf{x}^n)$ are drawn i.i.d. according to (5.15) *above.*

Proof. Using the bound on the $L_2(D,\varrho_D)$ error from Theorem 5.7 together with (7.2) above, we obtain (7.4) with high probability if the nodes $\mathbf{X} = (\mathbf{x}^1, ..., \mathbf{x}^n)$ are sampled according to the measure (5.15). □

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 probability measure and $\int_D K(\mathbf{x},\mathbf{x})\,\varrho_D(\mathrm{d}\mathbf{x}) < \infty$. 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 8.4.

When dealing with the problem of integrating an individual function, Theorem 6.1 has a direct consequence for Monte-Carlo integration.

Theorem 7.4. *Let μ_D be a measure on D such that $L_1(D,\varrho_D) \hookrightarrow L_1(D,\mu_D)$, where*

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

is the norm of the embedding. Drawing $(\mathbf{x}^1,...,\mathbf{x}^n)$ i.i.d. at random according to ϱ_D, we have under the assumptions of Theorem 6.1 for the conditional expectation of the individual integration error

$$\mathbb{E}\left(\left|\int_D f \,\mathrm{d}\mu_D - \mathrm{Q}^m_{\mathbf{X}} f\right|^2 \,\Big|\, \|\mathbf{H}_m - \mathbf{I}_m\| \le 1/2\right) \le \frac{C^2_{\varrho,\mu}}{1-\delta}\,\sigma_m^2\left(1+\frac{0.06}{\log n}\right).$$

Proof. Similar as in the proof of the previous theorem we estimate the integration error by the L_2-approximation error and apply Theorem 6.1 afterwards. □

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

Corollary 7.5. *Under the assumptions of Theorem 7.4, drawing $(\mathbf{x}^1,...,\mathbf{x}^n)$ i.i.d. at random according to ϱ_D, we have*

$$\mathbb{P}\left(\left|\int_D f \,\mathrm{d}\mu_D - \mathrm{Q}^m_{\mathbf{X}} f\right|^2 \le \frac{C^2_{\varrho,\mu}}{\delta}\,\sigma_m^2\left(1+\frac{0.06}{\log n}\right)\right) \ge 1-2\delta\,.$$

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 8.5) 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 [30] point out.

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

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

Defining the weight

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

and the univariate kernel function

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

directly leads to

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

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

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

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

$$f(\mathbf{x}) = \langle f, K^d_{\alpha,*}(\mathbf{x},\cdot)\rangle_{H^{\alpha,*}_{\text{mix}}}\,.$$

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

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

and the corresponding norm

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

The (ordered) sequence $(\lambda^*_j)_{j=1}^\infty$ of eigenvalues of the corresponding mapping $W=\mathrm{Id}^*\circ\mathrm{Id}$, where $\mathrm{Id}\colon H\big(K^d_{s,*}\big)\to L_2(\mathbb{T}^d)$ is the non-increasing rearrangement of the numbers

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

The corresponding orthonormal system $(e^*_j)_{j=1}^\infty$ in $H\big(K^d_{s,*}\big)$ is given by

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

Consequently, the orthonormal system $(\eta^*_j)_{j=1}^\infty$ in $L_2(\mathbb{T}^d)$ is the properly ordered classical Fourier system $\eta_{\mathbf{k}}(\mathbf{x}) = \exp(2\pi\mathrm{i}\,\mathbf{k}\cdot\mathbf{x})$. This directly implies the following behavior of the quantities $N(m)$ and $T(m)$ defined in (2.6) and (2.7), see also the comment after (2.8). 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 $\mathbf{s}=(s_1,...,s_d)$ to emphasize different smoothness in different coordinate directions. Such kernels will be denoted with $K_{\mathbf{s}}(\mathbf{x},\mathbf{y})$. In [35] the authors establish preasymptotic error bounds which can be used for the least squares analysis as we will see below.

Recent estimates in [33] 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

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

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

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

as well as the tensor product kernel

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

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

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

in the space $H^s_{\mathrm{mix}}(\mathbb{T}^d)$. 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 $W = \mathrm{Id}^*\circ\mathrm{Id}$, where $\mathrm{Id}\colon H\big(K^d_{s,\#}\big) \to L_2(\mathbb{T}^d)$, differs from $(\lambda^*_j)_{j=1}^\infty$ since the associated mappings $j\to\mathbf{k}$ do not coincide. Accordingly, the sampling operators $S^{m,*}_{\mathbf{X}}$ and $S^{m,\#}_{\mathbf{X}}$ 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 $H^s_{\mathrm{mix}}(\mathbb{T}^d)$, 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 [34, 33, 30]). Note that theoretically everything is known about the singular values $\sigma^{\square}_m$, $\square\in\{*,\#\}$, 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 started by Kühn, Sickel, T. Ullrich [34], where preasymptotic estimates of the type

$$\sigma^{\#}_m \le \left(\frac{\mathrm{e}^2}{m}\right)^{s/(2+\log_2(d))}\,, \quad 1\le m\le 4^d\cdot d/2\,, \tag{8.5}$$

are provided when considering the reproducing kernel $K^d_{s,\#}$. It was later noticed by Kühn [33] that for some special sequences, the restriction $1\le m\le 4^d\cdot d/2$ may be dropped and the exponent in (8.5) be improved. In more detail, the results in [33, Thm. 4.1] yield that the upper bound

$$\sigma^{\#}_m \le \min\left(1,\frac{16}{3m}\right)^{\frac{s}{1+\log_2 d}} \le \left(\frac{16}{3m}\right)^{\frac{s}{1+\log_2 d}} \tag{8.6}$$

holds for all $m\in\mathbb{N}$. One may argue that the kernel $K^d_{s,\#}$ is less "natural" than the kernel $K^d_{s,*}$. For this purpose we use the observations by Krieg [30]. It turned out that the size (measure) of the underlying domain D plays a significant role for the preasymptotic bound. In contrast to the result [34], where a version of the weight (8.2) has been considered on $D = [0, 2\pi]^d$, the preasymptotic behavior of the non-increasing rearrangement becomes much better when considering the modified problem on the unit cube $D = \mathbb{T}^d = [0, 1]^d$. To be more precise, Krieg obtained 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 linearly in s.

8.1 Recovery of individual functions from $H^s_{\text{mix}}(\mathbb{T}^d)$

In this section we aim at the recovery of a single individual function from random samples via least squares. Of course, this represents a less ambitious goal, since the random samples are not supposed to work uniformly for the whole class of functions. As one may expect this leads to better bounds. Note that, when doing numerical experiments with specific test functions, we actually observe the bounds for the individual recovery.

We have learned from our general Theorem 6.1 that we may relate the individual recovery error to the singular values / approximation numbers of the respective embedding. Together with Theorem 6.1 and the considerations in the previous section, this may be directly used to obtain preasymptotic error guarantees for the recovery of one single function using random samples. Note that this represents a special Monte-Carlo method. Theorem 8.2 below shows that our least squares approach has a slightly worse sampling complexity compared to the method in [31]. However, the numerical performance is striking and, in addition, it is robust with respect to Gaussian noise, which will be considered in a subsequent contribution.

Theorem 8.2. *Let $d \in \mathbb{N}, s > 1/2$ and $0 < \delta < 1$. Let further $n \in \mathbb{N}$ such that*

$$m := \left\lfloor \frac{n}{48(\sqrt{2}\log(2n) - \log \delta)} \right\rfloor + 1 \in [2, 3^d]\,.$$

Let f be a fixed function such that $\|f\|_* \leq 1$. We draw $\mathbf{X} = (\mathbf{x}^1, ..., \mathbf{x}^n)$ independently at random according to the uniform measure. Then the corresponding least squares operator $S^{m,*}_{\mathbf{X}}$ recovers f in the mean

$$\begin{aligned}\mathbb{E}\big(\|f - S^{m,*}_{\mathbf{X}} f\|_{L_2(\mathbb{T}^d)} \big| \|\mathbf{H}_m - \mathbf{I}_m\| \leq 1/2\big) \\ \leq \sqrt{\frac{1.06}{1-\delta}} \left(\frac{61(\sqrt{2}\log(2n) - \log \delta)}{n}\right)^{1.83s/(4+\log(d))}.\end{aligned}$$

With the help of Markov's inequality we also get a bound in probability.

Corollary 8.3. *Under the assumptions of Theorem 8.2, we obtain for a fixed function $\|f\|_* \leq 1$ that the recovery bound*

$$\|f - S^{m,*}_{\mathbf{X}} f\|_{L_2(\mathbb{T}^d)} \leq \sqrt{\frac{1.1}{\delta}} \left(\frac{61(\sqrt{2}\log(2n) - \log \delta)}{n}\right)^{1.83s/(4+\log(d))}$$

holds with probability exceeding $1 - 2\delta$.

Remark 8.4. When assuming that $\|f\|_\# \leq 1$, the corresponding operator $S_{\mathbf{X}}^{m,\#}$ recovers f up to the error

$$\|f - S_{\mathbf{X}}^{m,\#} f\|_{L_2(\mathbb{T}^d)} \leq \sqrt{\frac{1.1}{\delta}} \left(\frac{256(\sqrt{2}\log(2n) - \log\delta)}{n} \right)^{s/(1+\log_2(d))}$$

for all $n \in \mathbb{N}$ with probability exceeding $1 - 2\delta$. Here we use the bound in (8.6).

8.2 Uniform recovery of functions from $H^s_{\text{mix}}(\mathbb{T}^d)$

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 explicit, we study both cases collectively. For $\square \in \{*, \#\}$, the asymptotic behavior of the sequence $(\sigma_j^\square)_{j=1}^\infty$ for the embedding $\mathrm{Id} : H(K^d_{s,\square}) \to L_2(\mathbb{T}^d)$ has been known for a long time. There is a constant $\tilde{C}_d^\square$ which depends exponentially on d such that

$$\sigma_m^\square \leq \tilde{C}_d^\square\, m^{-s}(\log m)^{s(d-1)}\,, \quad m \in \mathbb{N}\,. \tag{8.7}$$

As a direct consequence of Corollary 5.6 we determine explicit error bounds as well as the asymptotic error behavior (8.8), where the latter holds for all equivalent norms in $H^s_{\text{mix}}(\mathbb{T}^d)$. Please note that the constants $C_d^\square$ depend heavily on the specific norm. Moreover, a statement similar to (8.8) can be also obtained with the technique in [32].

Theorem 8.5. *Let $d \in \mathbb{N}, s > 1/2$ and $n \in \mathbb{N}$ be given. Choose $m \in \mathbb{N}$, $m \geq 2$,*

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

Then, for $\delta > 0$ we achieve

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

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

$$\mathbb{P}\left(\sup_{\|f\|_\square \leq 1} \|f - S_{\mathbf{X}}^{m,\square} f\|_{L_2(\mathbb{T}^d)} \leq \frac{C_d^\square}{\delta} n^{-s}(\log n)^{sd} \right) \geq 1 - \delta\,. \tag{8.8}$$

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 $T_\square(m)$, cf. (2.7), using [11]

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

Hence, the right-hand side in (5.10) can be bounded from above by a constant times $\sigma_m^\square$, which behaves as $n^{-s}(\log n)^{sd}$. □

Remark 8.6. To our best knowledge, the bound in the previous theorem first showed up in Bohn [4, Sec. 5.5.4]. There the approximation error has been given for an individual function. We have already considered this problem in the last paragraph.

In addition, we investigate the preasymptotic error behavior using the aforementioned estimates (8.6) on the singular values $\sigma_m^{\#}$ that belongs to $\mathrm{Id}\colon H\big(K_{s,\#}^d\big) \to L_2(\mathbb{T}^d)$. 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 $w_s(\mathbf{k}) \ge \prod_{j=1}^d (1+|k_j|)^s$, which is fulfilled for $w_{s,\#}$.

Theorem 8.7. *Let $d \in \mathbb{N}$, $s > (1+\log_2 d)/2$, $0<\delta<1$, and $n \in \mathbb{N}$ such that*

$$m := \left\lfloor \frac{n}{48(\sqrt{2}\log(2n) - \log\delta)} \right\rfloor + 1 \tag{8.9}$$

is at least 2. Then, we obtain

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

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

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

for large enough n. Furthermore, we set $c := 16/3$, $\beta := 2s/(1+\log_2 d)$ and estimate $T_{\#}(m)$, cf. (2.7) and (8.6),

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

Taking (5.4) into account, we bound

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

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 $m \ge 2$ and thus, we need at least $n \ge 464$. This choice of n leads to an upper bound of the term within the brackets which is 29/6. Thus, for $m \ge 2$, the estimate

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

holds and the assertion follows. □

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 8.8. *Under the same assumptions as in Theorem 8.7 it holds*

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

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

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

□

Example 8.9. For $s = 5$ and $d = 16$, we would like to fulfill

$$\mathbb{P}\Bigg(\sup_{\|f\|_{\#}\le 1} \|f - S_{\mathbf{X}}^{m,\#} f\|_{L_2(\mathbb{T}^d)} \le 0.1 \Bigg) \ge 0.99\,,$$

i.e., with $\delta = 0.005$ we choose $m := 2\,346$ smallest possible such that

$$2\,\frac{29}{0.03} \left(\frac{16}{3m} \right)^2 \le 0.01$$

holds. Clearly, for n such that $m-1 = 2\,345 \le \frac{n}{48(\sqrt{2}\log(2n)-\log\delta)}$ holds, we observe the desired estimate. We choose the smallest possible $n := 3\,085\,266$.

8.3 Sampling recovery of multivariate functions from $H^s_{\text{mix}}(\mathbb{T}^d)$

Let us return to the problem of optimal sampling recovery of functions, described in Section 5.3. Here we consider the minimal worst-case error (sampling numbers/widths) defined by

$$g_n(\mathrm{Id} : H^s_{\text{mix}}(\mathbb{T}^d) \to L_2(\mathbb{T}^d)) := \inf_{\mathbf{X}=(\mathbf{x}^1,\ldots,\mathbf{x}^n)} \inf_{\substack{R:\,\mathbb{C}^n\to L_2 \\ \text{linear}}} \sup_{\|f\|_{H^s_{\text{mix}}}\le 1} \|f - R(f(\mathbf{X}))\|_{L_2(\mathbb{T}^d)}\,.$$

The correct asymptotic behavior of this quantity is one of the central open questions in Information Based Complexity (IBC), see e.g. [43], and Hyperbolic Cross Approximation, see [16, Outstanding Open Problem 1.4]. It has been shown by several authors, see e.g. [53] and [16, Sec. 5] for some historical remarks, that for $s > 1/2$ the bound

$$c_d n^{-s} (\log n)^{(d-1)s} \lesssim g_n(\mathrm{Id}) \lesssim C_d n^{-s} (\log n)^{(d-1)(s+1/2)} \tag{8.10}$$

holds asymptotically in $n \in \mathbb{N}$. Note that there is a gap in the logarithm between upper and lower bound. Recently, Krieg and M. Ullrich [32] improved this bound by using a probabilistic technique to show that

$$g_n(\mathrm{Id}) \lesssim_d n^{-s} (\log n)^{ds}\,.$$

Clearly, if $s < (d-1)/2$, then the gap in (8.10) is reduced to $(\log n)^s$. Let us emphasize that the result by Krieg, M. Ullrich [32] can be considered as a major progress for the research on the complexity of this problem. They disproved Conjecture 5.6.2. in [16] for $p = 2$ and $1/2 < s < (d-1)/2$. Indeed, the celebrated sparse grid points are now beaten by random points in a certain range for s. This again reflects the "power of random information", see [25]. 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. The problem is somehow related to the recent efforts in compressed sensing. There the optimal RIP matrices are given as realizations of random matrices. Known deterministic constructions are far from being optimal.

The outstanding open question remains (see e.g. [16], [43], and references therein) whether there is an intrinsic additional difficulty when restricting to algorithms based on function samples rather than Fourier coefficients. From a practical point of view, sampling algorithms are highly relevant since we usually have given discrete samples of functions. The question remains: are the asymptotic characteristics σ_n and g_n of the same order, or do they rather behave like $\sigma_n = o(g_n)$? This represents a fundamental open problem in hyperbolic cross approximation, see [16, Outstanding Open Problem 1.4]. Surprisingly, there is also recent progress in this direction. It turned out that the optimal sampling recovery problem is connected to the solution of the famous Kadison-Singer problem, see [38]. An interesting consequence of the groundbreaking result in [38] has been obtained by the authors in [40]. It reads as follows.

Theorem 8.10. *Let $(\eta_k)_{k=1}^{m'}$ be an orthonormal system in $L_2(D, \varrho_D)$ such that*

$$\frac{m'}{2} \leq \sum_{k=1}^{m'} |\eta_k(\mathbf{x})|^2 \leq 2m' .$$

Then there exists a set $\mathbf{X} = (\mathbf{x}^1, ..., \mathbf{x}^n) \in D^n$ with $n \leq C_1 m'$ such that for every $p(\mathbf{x}) = \sum\limits_{k=1}^{m'} c_k \eta_k(\mathbf{x})$ we have

$$C_2 \sqrt{\frac{1}{n} \sum_{j=1}^{n} |p(\mathbf{x}^j)|^2} \leq \|p\|_2 \leq C_3 \sqrt{\frac{1}{n} \sum_{j=1}^{n} |p(\mathbf{x}^j)|^2} .$$

This shows on the one hand that the problem of optimal recovery is strongly connected to the Marcinkiewicz discretization problem, see e.g. Gröchenig, Bass [1], the recent paper Gröchenig [23], and the references therein. Indeed, recently this family of problems gained renewed interest and has been studied systematically by Temlyakov and coauthors, see [57, 58, 18]. On the other hand, every d-variate trigonometric polynomial

$$t(\mathbf{x}) = \sum_{\mathbf{k} \in I} c_{\mathbf{k}} \exp(\mathrm{i}\, \mathbf{k} \cdot \mathbf{x})$$

with a fixed index set I, $|I| = m - 1$, may be stably recovered from $n = C|I|$ samples, since the matrix

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

has a condition number bounded by C_3/C_2 which, however, may depend badly on d. This is a surprising fact, since the well-known "$\log m$-oversampling" was out of question for a long time. From an algorithmic point of view, the construction above is hard to apply in practice since the existence of the point set is shown probabilistically without providing a small failure probability. However, this construction represents a promising candidate to make progress for the "Outstanding Open Problem" in [16]. In fact, this point construction behaves well in Proposition 5.3. The question remains open how to control the spectral norm of the infinite matrix in Proposition 5.4.

In this paper we also address algorithmic issues. The least squares regression (on random nodes) improves on the sparse grid technique with high probability. We also give a rather accurate bound on the above sampling numbers which are also reasonable for small n. Indeed, taking Theorem 8.5 with fixed $\delta = 29/64$ into account, we observe

$$g_n(\mathrm{Id}) \leq 8 \cdot \max\left\{\sigma_m, \sqrt{\frac{\log(8n)}{n}\sum_{j=m}^{\infty}\sigma_j^2}\right\},$$

where $m := \lfloor n\,(\sqrt{2}\log(2n)+1)^{-1}/48\rfloor + 1$ and $(\sigma_j)_{j\in\mathbb{N}}$ denotes the sequence of singular numbers of the embedding $\mathrm{Id} : H(K) \to L_2(\mathbb{T}^d)$, which are known also in a preasymptotic regime, see the previous subsection.

8.4 Numerical integration of periodic functions

In [45, Sec. 4.2], the author discussed the construction of stable cubature weights for the approximation of the integral

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

for functions from periodic Sobolev spaces with dominating mixed smoothness. The integration nodes $\mathbf{X}$ are drawn uniformly and independently at random from $[0,1]^d$. Below, in Corollary 8.11, the cubature rule $\mathrm{Q}_{\mathbf{X}}^m$ is fixed for the whole class. In contrast to that, we study a Monte-Carlo method in Theorem 8.13 below. The result in [45, 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. [45, Rem. 4.6]. Our theoretical results of this section confirm that the optimal main rate of the presented approach in [45] 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 8.11. *Let $d \in \mathbb{N}$, $s > 1/2$ and $0 < \delta < 1$. We choose $n \in \mathbb{N}$ such that m as stated in (8.9) is at least 2. Drawing $\mathbf{X} = (\mathbf{x}^1, \ldots, \mathbf{x}^n)$ uniformly i.i.d. at random from $[0,1]^d$ we put* *the cubature weight vector* $\boldsymbol{q}$ *to be the first column of* $\overline{\mathbf{L}_m}\,\overline{(\mathbf{L}_m^*\,\mathbf{L}_m)^{-1}}$. *Then, with probability at least* $1-2\delta$,

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

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

$$\sup_{\|f\|_\square\le 1}|I(f)-\mathrm{Q}_{\mathbf{X}}^m f|\le\frac{C_d^\square}{\delta}n^{-s}(\log n)^{sd}. \tag{8.11}$$

Proof. We apply Theorem 7.1 with $\mu_{\mathbb{T}^d}=\varrho_{\mathbb{T}^d}\equiv 1$ followed by (8.7). $\square$

Remark 8.12. By the same reasoning, the result in Theorem 8.7 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 [20, pp. 195, 247].

Now we consider the problem of integrating one particular function with random sampling points and a cubature formula that uses optimal weights, cf. [45]. Interestingly, we may reduce the general bound in Theorem 7.4 by a logarithmic factor in this specific situation. We stress the fact that we construct a cubature rule which uses sampling values from a set of random nodes similar to the Monte-Carlo method. On the one hand, computing the optimal weights causes additional computational effort compared to plain Monte-Carlo. On the other hand, this strategy yields substantially improved cubature rules.

Theorem 8.13. *Let* $d\in\mathbb{N}, s>1/2$ *and* $0<\delta<1$. *We choose* $n\in\mathbb{N}$ *such that* m *as stated in* (8.9) *is at least* 2. *Let further* f *be a fixed function such that* $\|f\|_{H^s_{\mathrm{mix}}(\mathbb{T}^d)}\le 1$. *Then we get*

$$\mathbb{E}\big(|\,\mathrm{I}(f)-\mathrm{Q}_{\mathbf{X}}^m f|^2\mid\|\mathbf{H}_m-\mathbf{I}_m\|\le 1/2\big)\le\frac{0.06}{\log(2n)}\sigma_m^2\,.$$

Proof. We first observe that

$$\mathrm{I}(f)-\mathrm{I}(S_{\mathbf{X}}^m f)=\mathrm{I}(P_{m-1}f)-\mathrm{I}(S_{\mathbf{X}}^m P_{m-1}f)+\mathrm{I}(g)-\mathrm{I}(S_{\mathbf{X}}^m g)$$

with $g=f-P_{m-1}f$. Due to the special structure of the Fourier system ($\mathrm{I}(g)=0$) and the reproduction property of $S_{\mathbf{X}}^m$ we find

$$|\,\mathrm{I}(f)-\mathrm{Q}_{\mathbf{X}}^m f)|^2=|\,\mathrm{I}(f)-\mathrm{I}(S_{\mathbf{X}}^m f)|^2=|\,\mathrm{I}(S_{\mathbf{X}}^m g)|^2\le\|S_{\mathbf{X}}^m g\|_{L_2(\mathbb{T}^d)}^2\,.$$

Examining the proof of Theorem 6.1 once again, we see that we have already estimated $\|S_{\mathbf{X}}^m g\|_{L_2(\mathbb{T}^d)}^2$. In (6.2) and what follows, we obtained

$$\mathbb{E}\big(\|S_{\mathbf{X}}^m g\|_{L_2(\mathbb{T}^d)}^2\big)\le\frac{4\,N(m)}{n}\sigma_m^2\le\frac{0.06}{\log(2n)}\sigma_m^2\,.$$

Hence, we find

$$\int_{\|\mathbf{H}_m-\mathbf{I}_m\|\le 1/2}|\,\mathrm{I}(f)-\mathrm{Q}_{\mathbf{X}}^m f|^2\,(\mathrm{d}\mathbf{x})^n\le\frac{0.06}{\log(2n)}\sigma_m^2\,.$$

$\square$

Making the right-hand side in Theorem precise we obtain:

Corollary 8.14. *Under the assumptions of Theorem 8.13 we estimate*

$$\mathbb{E}\Big(|\,\mathrm{I}(f) - \mathrm{Q}_{\mathbf{X}}^{m,\#} f| \;\Big|\; \|\mathbf{H}_{m,\#} - \mathbf{I}_m\| \le 1/2\Big) \le \frac{0.25}{\sqrt{\log(2n)}} \left(\frac{256(\sqrt{2}\log(2n) - \log\delta)}{n}\right)^{\frac{s}{1+\log_2 d}} .$$

9 Hyperbolic wavelet regression

The following scenario of replacing the Fourier system by dyadic wavelets has already been investigated by Bohn [4, Sec. 5.5.2], [5]. Using orthogonal wavelets we will improve the result in [4] in two directions. First, due to the vanishing moments of the wavelets, we will get a better error bound since the hyperbolic wavelet system is L_2-stable. 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 [19].

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

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

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

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

It is well-known, that $H^s_{\mathrm{mix}}(\mathbb{R}^d)$ 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

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

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 $f \in H^s_{\mathrm{mix}}(\mathbb{R}^d)$ 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

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

Clearly, if $\|f\|_{H^s_{\mathrm{mix}}(\mathbb{R}^d)} \le 1$, 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 $Q = [0,1]^d$. Let further $D_{\mathbf{j}}$ the set of all $\mathbf{k} \in \mathbb{Z}^d$ such that the wavelet $\operatorname{supp} \psi_{\mathbf{j},\mathbf{k}}$ has a non-empty intersection with $[0,1]^d$. This directly leads to the extended domain Ω given by

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

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 $f \in H^s_{\mathrm{mix}}(\mathbb{R}^d)$ 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 $\mathbf{X} = (\mathbf{x}^1, ..., \mathbf{x}^n)$ of the sampling nodes are chosen in advance for the whole class of functions. Let us consider the operator

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

which is known from *hyperbolic wavelet approximation*, see [19, 54]. The following worst-case error bound is well-known and follows directly from (9.1):

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

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

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

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

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 $S^\ell_{\mathbf{X}}$ described above recovers any $f \in H^s_{\mathrm{mix}}(\mathbb{R}^d)$ 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,*

$$\sup_{\|f\|_{H^s_{\mathrm{mix}}(\mathbb{R}^d)} \le 1} \|f - S^\ell_{\mathbf{X}} f\|_{L_2([0,1]^d)} \lesssim C_{\delta,d} 2^{-\ell s}$$

with probability larger than $1 - \delta$. The operator $S^\ell_{\mathbf{X}}$ uses $n(\ell) \asymp 2^\ell \ell^d$ many samples, which results in the complexity bound

$$\sup_{\|f\|_{H^s_{\mathrm{mix}}(\mathbb{R}^d)} \le 1} \|f - S^\ell_{\mathbf{X}} f\|_{L_2([0,1]^d)} \lesssim C_{\delta,d} n^{-s} \log(n)^{ds} . \tag{9.3}$$

Algorithm 3 Hyperbolic wavelet regression.

Input: $\ell \in \mathbb{N}$,

n: $\tilde{N}(\ell) \asymp 2^\ell \ell^{d-1} \lesssim \frac{n}{\log(n)}$,

$\mathbf{X} = (\mathbf{x}^1, ..., \mathbf{x}^n) \in D^n$ set of distinct sampling nodes,

$\mathbf{f} = (f(\mathbf{x}^1), ..., f(\mathbf{x}^n))^\top$ samples of f evaluated at the nodes from $\mathbf{X}$, 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 [47]), 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 $S^\ell_{\mathbf{X}} f := \sum\limits_{\|\mathbf{j}\|_1 \le \ell} \sum\limits_{\mathbf{k} \in D_{\mathbf{j}}} c_{\mathbf{j},\mathbf{k}} \, \widetilde{\psi}_{\mathbf{j},\mathbf{k}}$.

Remark 9.2. *(i)* Note that the optimal operator $\tilde{P}_\ell$ 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. [16], [41, 42, 44], 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 $\asymp \#\{\|\mathbf{j}\|_1 \le \ell\} \asymp \ell^d$ many non-zero entries. This is 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 $\mathbf{X} = (\mathbf{x}^1, ..., \mathbf{x}^n)$ be randomly drawn from Ω according to ϱ_Ω. If the number n of samples satisfies (5.1), that is

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

then

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

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^\ell_{\mathbf{X}} f\|_{L_2([0,1]^d)} &\le \|f - \tilde{P}_\ell f\|_{L_2([0,1]^d)} + \|\tilde{P}_\ell f - S^\ell_{\mathbf{X}} f\|_{L_2([0,1]^d)} \\
&\lesssim \Big\| f - \sum_{\|\mathbf{j}\|_1 \le \ell} \sum_{\mathbf{k}\in\mathbb{Z}^d} \langle f, \psi_{\mathbf{j},\mathbf{k}}\rangle \psi_{\mathbf{j},\mathbf{k}} \Big\|_{L_2(\mathbb{R}^d)} + \|S^\ell_{\mathbf{X}}(\tilde{P}_\ell f - f)\|_{L_2(\Omega,\varrho_\Omega)} \\
&\lesssim 2^{-\ell s} + \Big\| (\tilde{\mathbf{L}}_\ell^* \tilde{\mathbf{L}}_\ell)^{-1} \tilde{\mathbf{L}}_\ell^* \Big\| \cdot \left(\sum_{u=1}^n \Big| \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) \Big|^2 \right)^{1/2} \\
&\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 \Bigg[\sum_{\|\mathbf{j}'\|_1 > \ell} \sum_{\mathbf{k}'\in\mathbb{Z}^d} \sum_{\|\mathbf{j}\|_1 > \ell} \sum_{\mathbf{k}\in\mathbb{Z}^d} \\
&\qquad\qquad 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}} \Bigg]^{1/2} \\
&\lesssim 2^{-\ell s} + \sqrt{\frac{2}{n}} \big\| (2^{\|\mathbf{j}\|_1 s} \langle f, \psi_{\mathbf{j},\mathbf{k}}\rangle)_{\mathbf{j},\mathbf{k}} \big\|_2 \, \|\tilde{\mathbf{\Phi}}_\ell\| \\
&\lesssim 2^{-\ell s} + \sqrt{\frac{2}{n}} \|\tilde{\mathbf{\Phi}}_\ell\| \,,
\end{aligned}
$$

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

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

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

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

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

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

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

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

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

$$\mathbb{E}\|\tilde{\mathbf{\Phi}}_{\ell}\|^2 \lesssim n2^{-2\ell s}\,.$$

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. □

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 for periodic test functions f from the spaces $H^s_{\mathrm{mix}}(\mathbb{T}^d)$. 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 [28, 29] and [8]. This motivates us to consider Frolov lattices [27] and Fibonacci lattices (cf. e.g. [56, 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 (8.6), the weights without 2π lead to a slightly slower decay of the respective singular numbers.

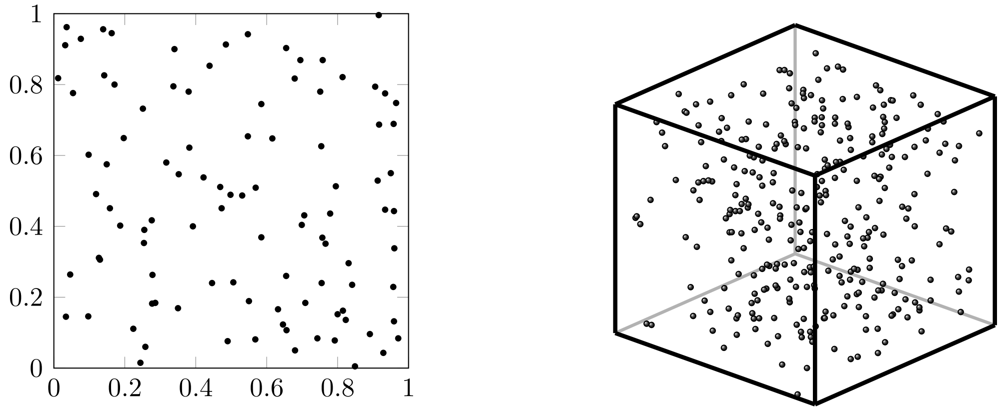
For a given number n of samples, we use the $m = \lfloor n/(4\log n)\rfloor$ frequencies $\mathbf{k} \in \mathbb{Z}^d$ where $w(\mathbf{k})$ is smallest. Here ties are broken in numerical order starting with the first component k_1 of $\mathbf{k}$ until the last one k_d. Corresponding to our theoretical results, the goal is to compute a least squares approximation $S^m_{\mathbf{X}}f$, $|\mathbf{X}| = n$, of the projection $P_{m-1}f$ of the function f to the span$\{\exp(2\pi\mathrm{i}\,\mathbf{k}\cdot\mathbf{x})\colon \mathbf{k} \in I\}$.

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

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

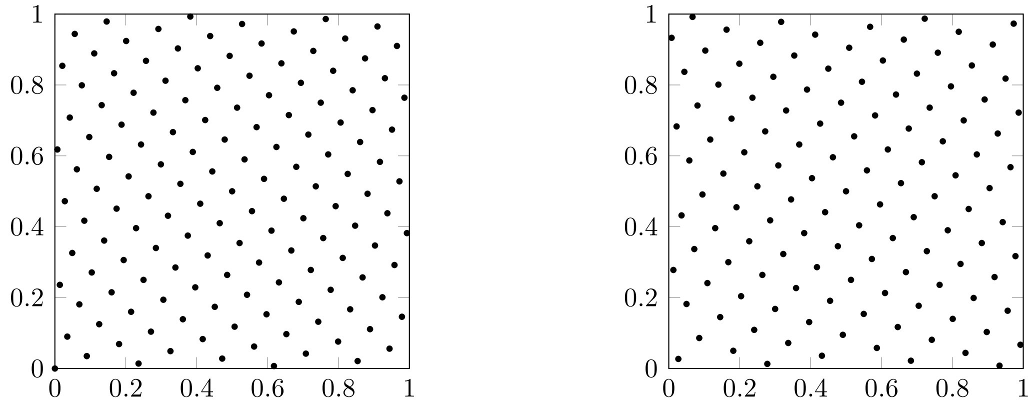
arithmetic operations, and setting up the matrix $\mathbf{L}_m$ requires

$$\leq C_2\,d\,n\,m$$



(a) $d = 2$, $n = 100$ (b) $d = 3$, $n = 316$

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



(a) Fibonacci lattice $n = 144$ (b) Frolov lattice $n = 127$

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

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 8.2 or 8.5, cf. (5.1), 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 $\approx 10^{-8}$. Here in our experiments, we choose $m = \lfloor n/(4 \log n) \rfloor$ slightly larger.

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 [28, 8], 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 8.5. 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.

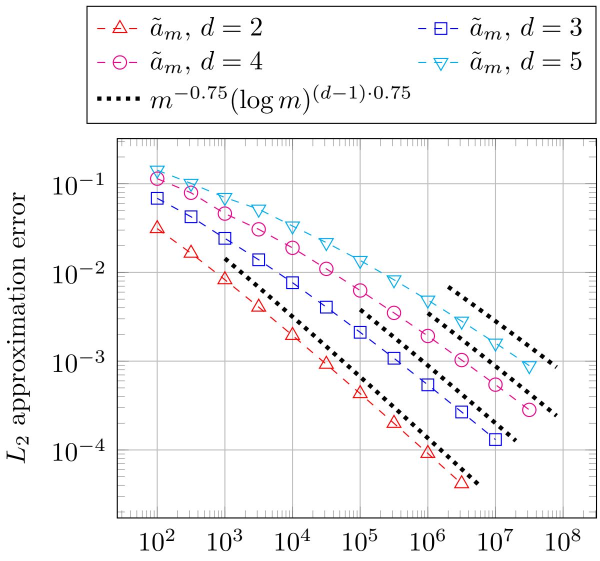
Subsequently, we consider three different test functions $f\colon \mathbb{T}^d \to \mathbb{R}$, where the Fourier coefficients $\hat{f}_{\mathbf{k}} := \int_{\mathbb{T}^d} f(\mathbf{x}) \exp(-2\pi\mathrm{i}\,\mathbf{k}\cdot\mathbf{x})\,\mathrm{d}\mathbf{x}$, $\mathbf{k} \in \mathbb{Z}^d$, of f, decay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

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

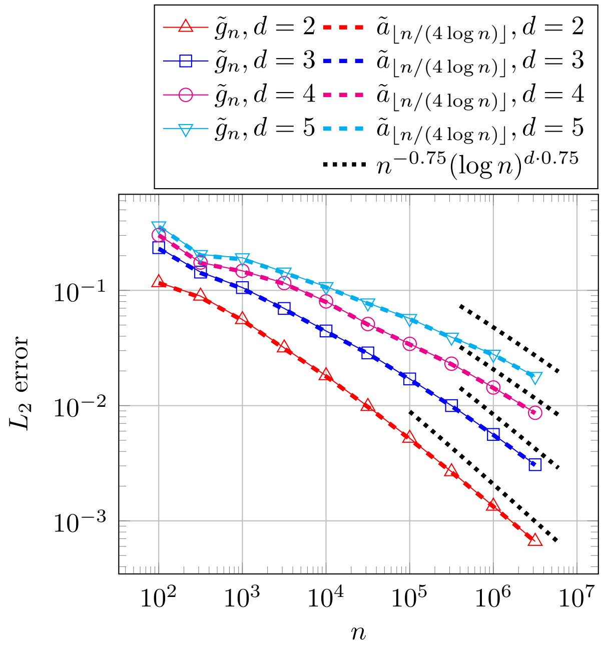
We start with the test function

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

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



(a) approximation error

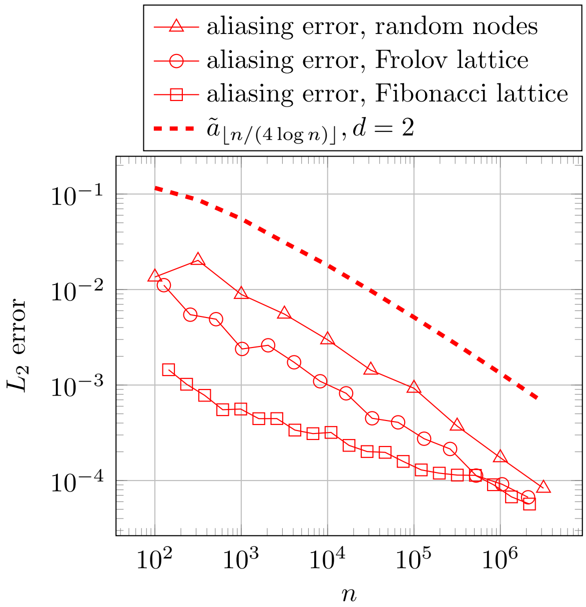


(b) sampling error

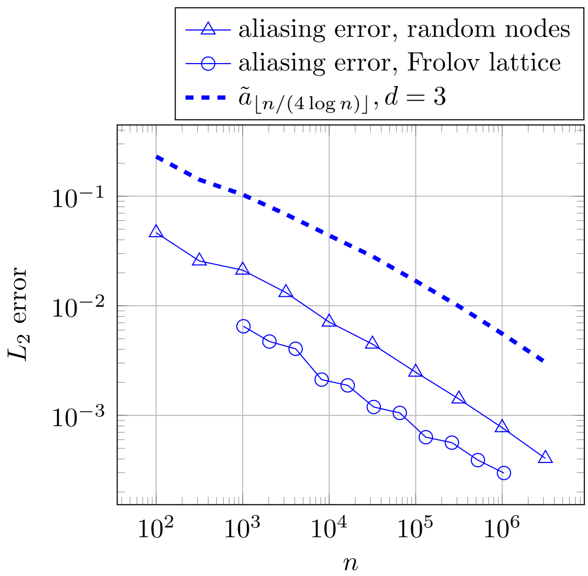
Figure 3: Approximation errors and least squares sampling errors for test function $f \in H^{3/4-\varepsilon}_{\mathrm{mix}}(\mathbb{T}^d)$.

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 (8.7), 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 [47] on the matrix $\mathbf{L}_m$, $m = \lfloor n/(4\log n)\rfloor$. The obtained sampling errors $\tilde{g}_n := \|f - S^m_{\mathbf{X}} f\|_{L_2(\mathbb{T}^d)}$ are visualized in Figure 3b as well as the graphs $\sim n^{-0.75}(\log n)^{d\cdot 0.75}$ as dotted lines which correspond to the theoretical upper bounds $n^{-0.75+\varepsilon}(\log n)^{d\cdot(0.75-\varepsilon)}$, cf. (8.8). 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.



(a) $d=2$



(b) $d=3$

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

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 $\|P_{m-1}f-S^m_{\mathbf{X}}f\|_{L_2(\mathbb{T}^d)}$, $m=\lfloor n/(4\log n)\rfloor$, which are the errors caused by Algorithm 1 since

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

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, so-called Frolov lattices [27] 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. They decay similarly and are lower than the aliasing errors for random nodes in most cases.

In addition, we consider Fibonacci lattices. 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 $S^m_{\mathbf{X}}f$ 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 $H^{3/2-\varepsilon}_{\text{mix}}(\mathbb{T}^d)$

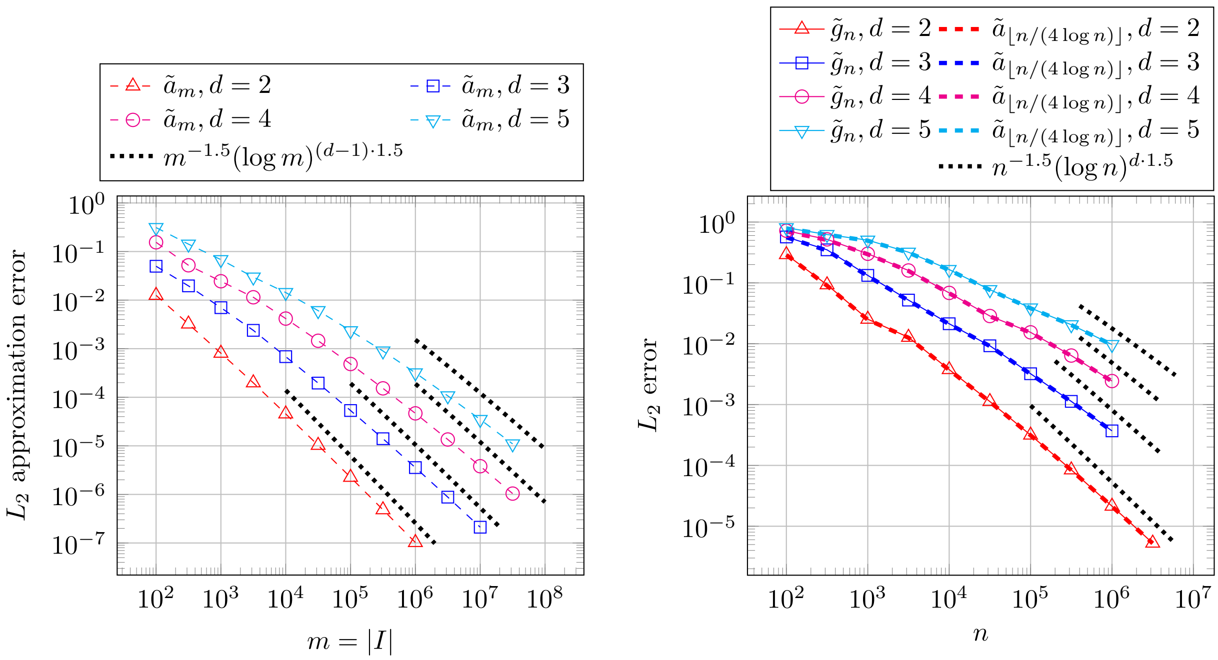
Next, we consider the kink test function $f\colon \mathbb{T}^d \to \mathbb{R}$,

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

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.



(a) approximation error

(b) sampling error

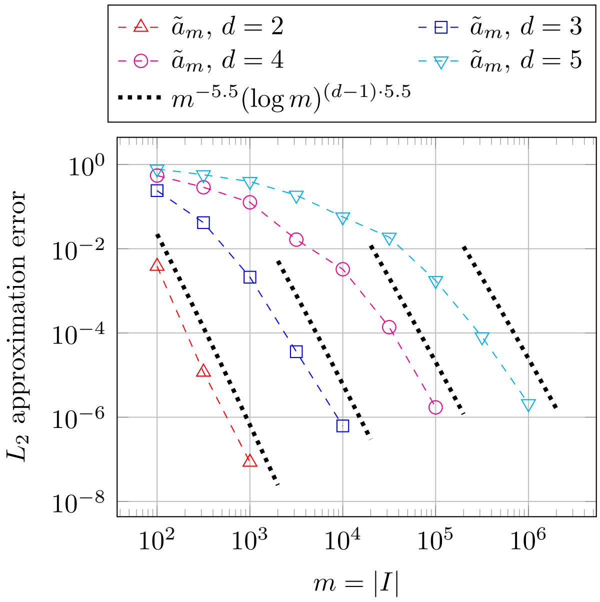
Figure 5: Approximation errors and least squares sampling errors for test function $f \in H^{3/2-\varepsilon}_{\text{mix}}(\mathbb{T}^d)$.

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 $\tilde{g}_n := \|f - S^m_{\mathbf{X}} f\|_{L_2(\mathbb{T}^d)}$ are depicted in Figure 5b. In addition, the graphs $\sim n^{-1.5}(\log n)^{d \cdot 1.5}$ 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 $m = \lfloor n/(4\log n)\rfloor$

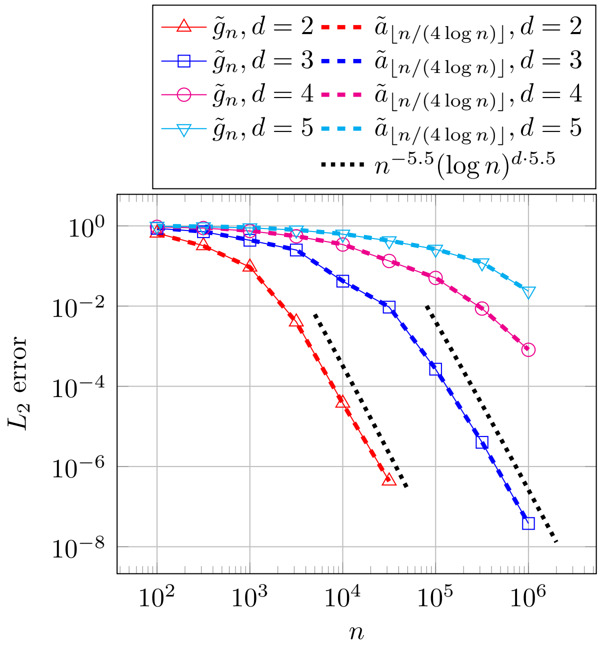
as thick dashed lines, and we observe that these approximation errors $\tilde{a}_{\lfloor n/(4\log n)\rfloor}$ almost coincide with the sampling errors. Correspondingly, we still have preasymptotic behavior for $d = 3, 4, 5$.

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

As a third test function f, we consider a periodic B-Spline of order 6, which is a piecewise polynomial of degree 5, and therefore, we have $f \in H^{11/2-\varepsilon}_{\text{mix}}(\mathbb{T}^d)$.



(a) approximation error



(b) sampling error

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

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 $\tilde{g}_n := \|f - S^m_{\mathbf{X}} f\|_{L_2(\mathbb{T}^d)}$ 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 $\sim n^{-5.5}(\log n)^{d\cdot 5.5}$ 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 [45], cf. in particular [45, Sec. 4.1] for tests concerning numerical integration in Sobolev spaces with mixed smoothness $H^s_{\text{mix}}(\mathbb{T}^d)$. 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 [45, 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 8.11, i.e., our theoretical main rate corresponds to the observations by Oettershagen [45]. Moreover, we have an error bound of $\sim n^{-s}(\log n)^{sd-1/2}$ for an individual function $f \in H^s_{\mathrm{mix}}(\mathbb{T}^d)$ as well as preasymptotic results, cf. Theorem 8.13 and Corollary 8.14.

10.3 Recovery and integration for the non-periodic case

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

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

and its tensor product

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

The space $\tilde{H}^s_{\mathrm{mix}}([-1,1]^d)$ 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, $k=1,\ldots,m-1$, by $\boldsymbol{h}_k := \left(h_{k,t}\right)_{t=1}^d \in \mathbb{N}_0^d$, 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 $\boldsymbol{b}\in\mathbb{C}^{m-1}$ in (7.1), we have

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

for $\mu_D\equiv 1$.

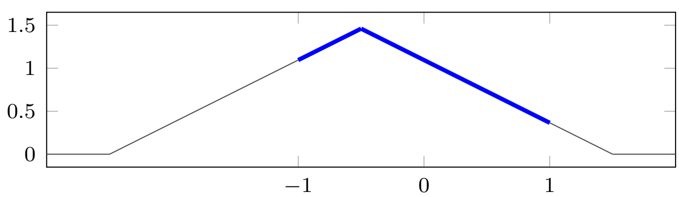


Figure 7: Dilated, scaled, and shifted B-Spline of order 2 considered in interval $[-1,1]$.

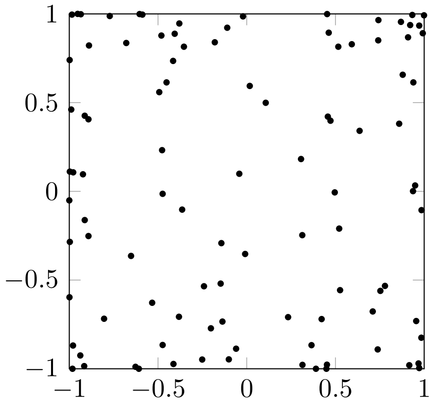
For the numerical experiments we consider the test function

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

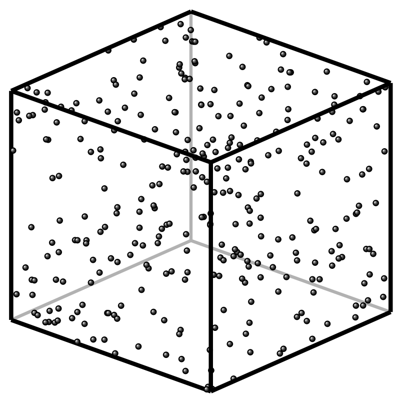
on $D := [-1,1]^d$, where the one-dimensional version is depicted in Figure 7. For the Chebyshev coefficients of f, we have

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

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

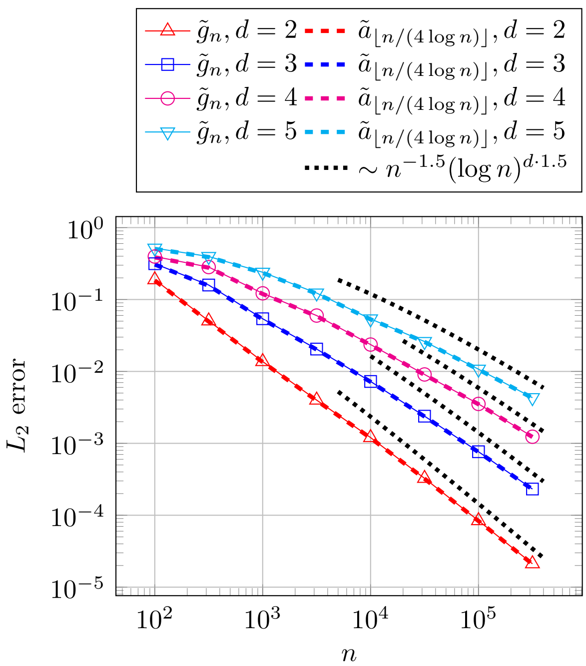


(a) $d = 2$, $n = 100$

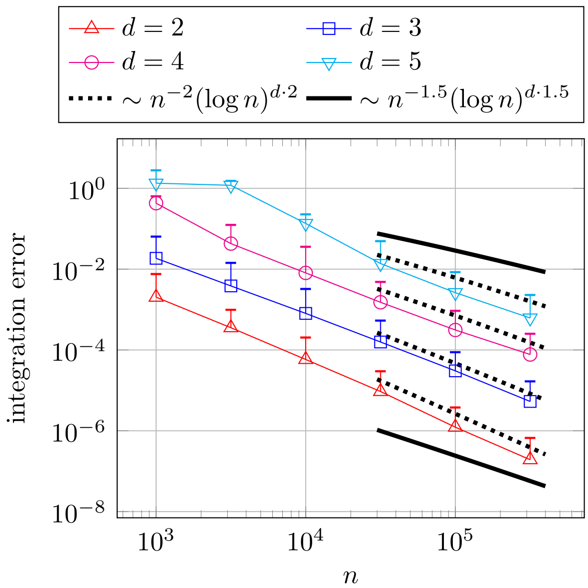


(b) $d = 3$, $n = 316$

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



(a) sampling and approximation eror



(b) integration error

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

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

$\tilde{\varrho}_D(\mathrm{d}\mathbf{x}) := \prod_{i=1}^d (\pi\sqrt{1-x_i^2})^{-1}\,\mathrm{d}\mathbf{x}$. 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 $\tilde{g}_n := \|f - S_{\mathbf{X}}^m f\|_{L_2([-1,1]^d,\varrho_D)}$ with $m = \lfloor n/(4\log n)\rfloor$ are plotted in Figure 9a as triangles as well as the corresponding approximation errors $\tilde{a}_m := \|f - P_{m-1}f\|_{L_2([-1,1]^d,\varrho_D)}$ as thick dashed lines. We observe that the sampling and approximation errors almost coincide. Moreover, we plot the graphs $\sim n^{-1.5}(\log n)^{d\cdot 1.5}$ as dotted lines which correspond to the expected theoretical upper bounds $n^{-1.5+\varepsilon}(\log n)^{d\cdot(1.5-\varepsilon)}$. 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 $Q_{\mathbf{X}}f = \int_D S_{\mathbf{X}}^m f\,\mathrm{d}\mu_D = \sum_{k=1}^m c_k\, b_k$ 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 $\sim n^{-2}(\log n)^{d\cdot 2}$ as dotted lines, and we observe that the obtained integration errors approximately decay like these graphs. For comparison, we also plot $n^{-1.5}(\log n)^{d\cdot 1.5}$ for $d = 2$ and $d = 5$ as thick solid lines which belong to the theoretical results $n^{-1.5+\varepsilon}(\log n)^{d\cdot(1.5-\varepsilon)}$ one obtains analogously to (8.11) 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), (8.6) and [30] also hold for the Chebyshev case.

Acknowledgment. The authors would like to thank Bastian Bohn, Michael Griebel, Karlheinz Gröchenig, Aicke Hinrichs, Boris S. Kashin, David Krieg, Thomas Kühn, 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. *to appear, 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada., arXiv:1906.07832*, 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] B. Bohn. *Error analysis of regularized and unregularized least-squares regression on discretized function spaces.* Dissertation, Institut für Numerische Simulation, Universität Bonn, 2017.

[5] B. Bohn. On the convergence rate of sparse grid least squares regression. In J. Garcke, D. Pflger, 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.

[6] B. Bohn and M. Griebel. Error estimates for multivariate regression on discretized function spaces. *SIAM J. Numer. Anal.*, 55(4):1843–1866, 2017.

[7] G. Byrenheid. *Sparse representation of multivariate functions based on discrete point evaluations.* Dissertation, Institut für Numerische Simulation, Universität Bonn, 2018.

[8] 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.

[9] 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.

[10] A. Christmann and I. Steinwart. *Support Vector Machines.* Springer, 2008.

[11] 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.

[12] 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.

[13] 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.

[14] A. Cohen and G. Migliorati. Optimal weighted least-squares methods. *SMAI J. Comput. Math.*, 3:181–203, 2017.

[15] D. Dũng. B-spline quasi-interpolant representations and sampling recovery of functions with mixed smoothness. *J. Complexity*, 27(6):541–567, 2011.

[16] D. Dũng, V. N. Temlyakov, and T. Ullrich. *Hyperbolic Cross Approximation.* Advanced Courses in Mathematics. CRM Barcelona. Birkhäuser/Springer, 2019.

[17] D. Dũng 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.

[18] 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.

[19] R. A. DeVore, S. V. Konyagin, and V. N. Temlyakov. Hyperbolic wavelet approximation. *Constr. Approx.*, 14(1):1–26, 1998.

[20] J. Dick, F. Y. Kuo, and I. H. Sloan. High-dimensional integration: the quasi-Monte Carlo way. *Acta Numer.*, 22:133–288, 2013.

[21] S. Foucart and H. Rauhut. *A mathematical introduction to compressive sensing.* Applied and Numerical Harmonic Analysis. Birkhäuser/Springer, New York, 2013.

[22] K. Gröchenig. Irregular sampling, Toeplitz matrices, and the approximation of entire functions of exponential type. *Math. Comp.*, 68(226):749–765, 1999.

[23] K. Gröchenig. Sampling, Marcinkiewicz-Zygmund inequalities, approximation, and quadrature rules. *arXiv:1909.07752*, 2019.

[24] A. Hinrichs. Optimal importance sampling for the approximation of integrals. *J. Complexity*, 26(2):125–134, 2010.

[25] A. Hinrichs, E. Novak, D. Krieg, J. Prochno, and M. Ullrich. On the power of random information. *arXiv:1903.00681*, 2019.

[26] A. Hinrichs, E. Novak, and J. Vybíral. Linear information versus function evaluations for L_2-approximation. *J. Approx. Theory*, 153(1):97–107, 2008.

[27] C. Kacwin, J. Oettershagen, M. Ullrich, and T. Ullrich. Numerical performance of optimized Frolov lattices in tensor product reproducing kernel Sobolev spaces. *arXiv:math/1802.08666*, pages 1–40, 2018.

[28] 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.

[29] 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.

[30] D. Krieg. Tensor power sequences and the approximation of tensor product operators. *J. Complexity*, 44:30–51, 2018.

[31] D. Krieg. Optimal Monte Carlo methods for L_2-approximation. *Constr. Approx.*, 49(2):385–403, 2019.

[32] D. Krieg and M. Ullrich. Function values are enough for l_2-approximation. *arXiv:math/1905.02516v3*, 2019.

[33] T. Kühn. New preasymptotic estimates for the approximation of periodic Sobolev functions. In *2018 MATRIX annals*, volume 3 of *MATRIX Book Ser.* Springer, Cham, to appear, https://www.matrix-inst.org.au/2018-matrix-annals/.

[34] 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.

[35] T. Kühn, W. Sickel, and T. Ullrich. How anisotropic mixed smoothness affects the decay of singular numbers of Sobolev embeddings. *arXiv:2001.09022*, 2020.

[36] 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.

[37] F. Lust-Piquard and G. Pisier. Noncommutative Khintchine and Paley inequalities. *Ark. Mat.*, 29(2):241–260, 1991.

[38] 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.

[39] 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.

[40] S. Nitzan, A. Olevskii, and A. Ulanovskii. Exponential frames on unbounded sets. *Proc. Amer. Math. Soc.*, 144(1):109–118, 2016.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] J. Oettershagen. *Construction of Optimal Cubature Algorithms with Applications to Econometrics and Uncertainty Quantification.* Dissertation, Institut für Numerische Simulation, Universität Bonn, 2017.

[46] R. I. Oliveira. Sums of random Hermitian matrices and an inequality by Rudelson. *Electr. Comm. Probab.*, 15:203–212, 2010.

[47] 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.

[48] A. Papageorgiou and H. Woźniakowski. Tractability through increasing smoothness. *J. Complexity*, 26(5):409–421, 2010.

[49] 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.

[50] H. Rauhut and R. Ward. Sparse Legendre expansions via ℓ_1-minimization. *J. Approx. Theory*, 164(5):517–533, 2012.

[51] H. Rauhut and R. Ward. Interpolation via weighted ℓ_1 minimization. *Appl. Comput. Harmon. Anal.*, 40(2):321–351, 2016.

[52] M. Rudelson. Random vectors in the isotropic position. *J. Funct. Anal.*, 164(1):60–72, 1999.

[53] W. Sickel and T. Ullrich. The Smolyak algorithm, sampling on sparse grids and function spaces of dominating mixed smoothness. *East J. Approx.*, 13(4):387–425, 2007.

[54] 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.

[55] 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.

[56] V. N. Temlyakov. *Approximation of periodic functions*. Computational Mathematics and Analysis Series. Nova Science Publishers Inc., Commack, NY, 1993.

[57] V. N. Temlyakov. The Marcinkiewicz-type discretization theorems. *Constr. Approx.*, 48(2):337–369, 2018.

[58] V. N. Temlyakov. Sampling discretization error of integral norms for function classes. *J. Complexity*, 54:101408, 16, 2019.

[59] 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.

[60] H. Wendland. *Scattered data approximation*, volume 17 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 2005.