Error Guarantees for Least Squares Approximation with Noisy Samples in Domain Adaptation

Felix Bartel [1]

[1] Chemnitz University of Technology, Faculty of Mathematics, 09107 Chemnitz, Germany
Email address: felix.bartel@mathematik.tu-chemnitz.de.

Abstract. Given n samples of a function $f\colon D \to \mathbb{C}$ in random points drawn with respect to a measure ϱ_S we develop theoretical analysis of the $L_2(D, \varrho_T)$-approximation error. For a parituclar choice of ϱ_S depending on ϱ_T, it is known that the weighted least squares method from finite dimensional function spaces V_m, $\dim(V_m) = m < \infty$ has the same error as the best approximation in V_m up to a multiplicative constant when given exact samples with logarithmic oversampling. If the source measure ϱ_S and the target measure ϱ_T differ we are in the domain adaptation setting, a subfield of transfer learning. We model the resulting deterioration of the error in our bounds.

Further, for noisy samples, our bounds describe the bias-variance trade off depending on the dimension m of the approximation space V_m. All results hold with high probability.

For demonstration, we consider functions defined on the d-dimensional cube given in unifom random samples. We analyze polynomials, the half-period cosine, and a bounded orthonormal basis of the non-periodic Sobolev space H^2_{mix}. Overcoming numerical issues of this H^2_{mix} basis, this gives a novel stable approximation method with quadratic error decay. Numerical experiments indicate the applicability of our results.

Keywords. domain adaptation, individual function approximation, least squares, sampling theory, transfer learning, unit cube, polynomial approximation.

2020 Mathematics Subject Classification. 41A10, 41A25, 41A60, 41A63, 42C10, 65Txx, 65F22, 65D15, 94A20 .

1. Introduction

In this paper we study the reconstruction of complex-valued functions on a d-dimensional domain $D \subset \mathbb{R}^d$ from possibly noisy function values

$$\boldsymbol{y} = \boldsymbol{f} + \boldsymbol{\varepsilon} = (f(\boldsymbol{x}^1) + \varepsilon_1, \dots, f(\boldsymbol{x}^n) + \varepsilon_n)^{\mathsf{T}},$$

which are sampled in random points $\boldsymbol{x}^1, \dots, \boldsymbol{x}^n \in D$. We consider error bounds for the weighted least squares method for individual functions, which is common in, e.g. partial differential equations [9] or uncertainty quantification [22]. In this setting, the samples are drawn after the function is fixed in contrast to worst-case or minmax-bounds, which hold for a class of functions and usually do not include noise in the samples. For individual function approximation the majority of L_2-error bounds are stated in expectation, cf. [5, Thm. 1.1] for penalized least-squares, [10, Thm. 3] for plain least-squares or, [21, Thm. 4.1], and [24, Thm. 6.1] for weighted least squares. Bounds, which hold with high probability, are known for polynomial approximation, cf. [31, Thm. 3], wavelet approximation, cf. [29, Thms. 3.20 & 3.21], or in a more general setting incuding noise in [11, Thm. 4.3] with the coarser L_∞-norm instead of the natural L_2-norm in the estimate. Further, in [11, Thm. 4.1] an error bound with the natural L_2-norm estimate is presented in expectation with the same behaviour as we will present with high probability. The contribution and novelty of this paper is twofold:

- We use concentration inequalities to show error bounds in the L_2- and L_∞-norm which hold with high probability, including the noisy case. The behaviour of our bound is similar to

The author was supported by the Deutscher Akademischer Austauschdienst (DAAD).

[11, Thm. 4.1], which is stated in expectation. Approximating from an m-dimensional function space we achieve the best error up to a multiplicative constant using logarithmic oversampling. Note, there exists a distribution such that linear oversampling achieves the optimal error but this is not constructive, cf. [14]. Including noise, our bounds reflect the typical bias-variance trade off which one wants to balance to prevent over- or underfitting. The results enable to give performance guarantees for model selection strategies like the balancing principle [37, 30] or cross-validation [7, 6].

- For an application we have a look at approximation on the d-dimensional unit cube $[0,1]^d$ when samples are distributed uniform according to the Lebesgue measure. A result with focus on polynomial approximation in the one-dimensional space is [31, Thm. 3] which is improved by the general result [11, Thm. 2.1]. There, the aproximation error is estimated by the L_∞-error of the projection with high probability and to the more natural L_2-error of the projection in expectation. We obtain a bound by the L_2-error of the projection which also holds with high probability. A drawback of polynomials is the need for quadratic oversampling, which we show for the Legendre polynomials but holds in general, cf. [31]. To circumvent this, we use the eigenfunctions of the embedding $\mathrm{Id}\colon H^s \to L_2$ from the Sobolev space H^s for $s = 1, 2$ which allow for logarithmic oversampling. The H^1 basis, also known as half-period cosine, was introduced in [25] and has become the standard in many applications and is researched thouroughly, cf. [23, 53, 1, 2, 13, 46, 12, 27]. But also for functions in Sobolev spaces H^s of higher smoothness their convergence is limited to be linear in theory (the rate $3/2$ can be observed in practice). This can be improved by using the H^2 basis, examined theoretically in [3, Section 3] to have quadratic convergence. So far it is not used as it is prone to numerical errors and unusable for higher degree approximation. Here, we propose an approximation and prove its accuracy which leads to a numerically stable way for approximating non-periodic uniform data with quadratic convergence.

For a more detailed formulation we need some notation. Given an m-dimensional function space $V_m \subset L_2$, we define the best possible approximation (projection) to $f\colon D \to \mathbb{C}$ in V_m and its error:

$$P(f, V_m, L_p) = \operatorname*{arg\,min}_{g \in V_m} \|f - g\|_{L_p} \text{ and } e(f, V_m, L_p)_{L_q} = \|f - P(f, V_m, L_p)\|_{L_q}$$

for $p, q \in \{2, \infty\}$. Note, since V_m is finite-dimensional the minimum is actually attained. Following [5, 10, 31, 11, 24, 29], we use weighted least squares S_m, defined in (3.1), as underlying approximation method. Because of its linearity, the approximation error $\|f - S_m \boldsymbol{y}\|_{L_2}$ splits as follows:

$$\begin{aligned}\|f - S_m \boldsymbol{y}\|_{L_2}^2 &= e(f, V_m, L_2)_{L_2}^2 + \|P(f, V_m, L_2) - S_m \boldsymbol{y}\|_{L_2}^2 \\ &\le \underbrace{e(f, V_m, L_2)_{L_2}^2}_{\text{truncation error}} + 2\underbrace{\|P(f, V_m, L_2) - S_m \boldsymbol{f}\|_{L_2}^2}_{\text{discretization error}} + 2\underbrace{\|S_m \boldsymbol{\varepsilon}\|_{L_2}^2}_{\text{noise error}}.\end{aligned}$$

For fixed number of points n, we have a look at the behaviour with respect to m, the dimension of the approximation space V_m. The truncation error is the best possible benchmark and usually has polynomial decay m^{-s} for some rate $s \ge 1$ depending on f and the choice of V_m. We show, that the discretization error obeys the same decay as the truncation error. Thus, given logarithmic oversampling, we obtain the best possible error up to a multiplicative constant in the noiseless case, cf. Theorem 3.2.

Including noise, we show that we get an additional summand growing linear in m, cf. Theorems 1.1. This resambles the well-known bias-variance trade off modeling the over- and undersmoothing effects which one wants to balance, cf. [20, 37]. This linear behaviour in m is approved by [30, Thm. 4.9] (by using the regularization $g_\lambda(\sigma) = 1/(\lambda + \sigma)$ with $\lambda = 0$). An example of that behaviour for $D = [0, 1]$

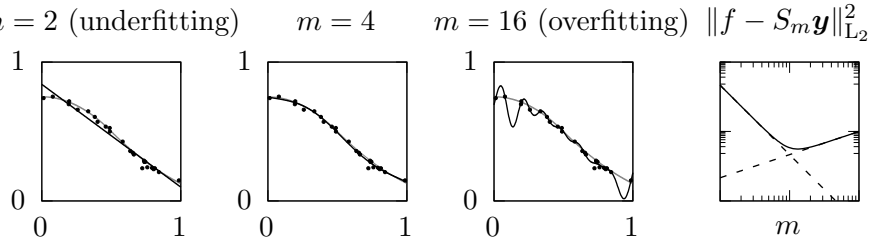

FIGURE 1.1. One-dimensional approximation on the unit-interval for three different choices of m and the schematic behaviour of the L_2-approximation error $\|f - S_m \boldsymbol{y}\|_{L_2}^2$ (solid line) split into the error for exact function values $\|f - S_m \boldsymbol{f}\|_{L_2}^2$ and the noise error $\|S_m \boldsymbol{\varepsilon}\|_{L_2}^2$ (dashed lines) with respect to m.

and $\varrho_T = \mathrm{d}x$ being the Lebesgue measure is depicted in Figure 4.1 where the detailed example is found in Section 4. Our central theorem, complying this behaviour, looks as follows:

Theorem 1.1. *Let $f\colon D \to \mathbb{C}$, $\boldsymbol{x}^1, \dots, \boldsymbol{x}^n$, $n \in \mathbb{N}$ be points drawn according to a probability measure $\mathrm{d}\varrho_S = 1/\beta \, \mathrm{d}\varrho_T$ and $\boldsymbol{y} = \boldsymbol{f} + \boldsymbol{\varepsilon} = (f(\boldsymbol{x}^1) + \varepsilon_1, \dots, f(\boldsymbol{x}^n) + \varepsilon_n)^{\mathsf{T}}$ noisy function values where $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^{\mathsf{T}}$ is a vector of independent complex-valued mean-zero random variables satisfying $\mathbb{E}(|\varepsilon_i|^2) \le \sigma^2$ and $|\varepsilon_i| \le B$ for $i = 1, \dots, n$. Let further, $t \ge 0$, V_m be an m-dimensional function space with an $L_2(D, \varrho_T)$-orthonormal basis $\eta_0, \dots, \eta_{m-1}$ satisfying*

$$10\|\beta(\cdot)N(V_m, \cdot)\|_\infty(\log(m) + t) \le n \quad \textit{with} \quad N(V_m, \cdot) = \sum_{k=0}^{m-1} |\eta_k(\cdot)|^2 .$$

Then, for S_m the weighted least squares method defined in (3.1) *with $\omega_i = \beta(\boldsymbol{x}^i)$, we have with joint probability exceeding $1 - 3\exp(-t)$:*

$$\begin{aligned}\|f - S_m \boldsymbol{y}\|_{L_2}^2 \le 14\Big(e(f, V_m, L_2)_{L_2} + \sqrt{\frac{t}{n}} e(f, V_m, L_2)_{L_\infty}\Big)^2 \\ + 4\|\beta\|_\infty \Big(\frac{m}{n}\Big(14B\sqrt{t\sigma^2} + \sigma^2\Big) + \frac{128B^2 t}{n}\Big) ,\end{aligned}$$

where $L_2 = L_2(D, \varrho_T)$ and $L_\infty = L_\infty(D, \varrho_T)$.

The first line of the bound corresponds to the truncation error and discretization error, decaying in m. Note, that the L_∞-term with the prefactor $n^{-1/2}$ behaves as the L_2-term whenever β is bounded from below, cf. Theorem 3.2. The second line is the error due to noise, increasing in m, cf. Figure 1.1. The estimation of the noise error is using a Hanson-Wright concentration inequality, which can be found using different assumptions. Thus, we can replace the noise model by general Bernstein conditions, cf. Lemma 2.3, or sub-Gaussian noise, cf. [43]. This theorem extends to the L_∞ case:

Theorem 1.2 (L_∞-error bound with noise)**.** *Let the assumptions of Theorem 1.1 hold. Then, for S_m the weighted least squares method defined in* (3.1) *with $\omega_i = \beta(\boldsymbol{x}^i)$, we have with probability exceeding $1 - 3\exp(-t)$:*

$$\begin{aligned}\|f - S_m \boldsymbol{y}\|_{L_\infty} \le \Big(1 + \sqrt{5N(V_m)}\Big)\Big(e(f, V_m, L_\infty)_{L_\infty} + \sqrt{\frac{t}{n}} e(f, V_m, L_\infty)_{L_2}\Big) \\ + 2\sqrt{\|\beta\|_\infty N(V_m)}\sqrt{\frac{m}{n}\Big(14B\sqrt{t\sigma^2} + \sigma^2\Big) + \frac{128B^2 t}{n}} .\end{aligned}$$

The bound is similar to [29, Thm. 3.21] in the wavelet setting but we use the best approximation with respect to the more natural L_∞ instead of L_2. In addition to the error of the best approximation we now have the additional factor $N(V_m)$ due to using the norm estimate $\|g\|_{L_\infty} \leq \sqrt{N(V_m)}\|g\|_{L_2}$ for functions $g \in V_m$. The same factor appears when approximating the worst-case error where it is known to be necessary in various examples, e.g. [41, Sec. 7] or [47, Thm .1.1].

The sampling measure $\varrho_S(E) \coloneqq \int_E 1/\beta \, d\varrho_T$, induced by the probability distribution β, may differ from the error measure ϱ_T, which is known as the *change of measure* and has applications in domain adaptation, cf. [36]. We assume to know β exactly but it may be approximated as well, cf. [18]. Note, β affects the maximal size of V_m in the assumption and the amplification of the noise in bound. There are two extremal cases:

(i) Having $\beta(\boldsymbol{x}) = m/N(V_m, \boldsymbol{x})$, as it was done in [22, 34, 11, 24], we obtain the assumption

$$10\|\beta(\cdot)N(V_m,\cdot)\|_\infty(\log(m)+t) = 10m(\log(m)+t) \leq n\,,$$

which allows for the biggest choice of m possible. But this spoils $\|\beta\|_\infty$ in the error bound when the Christoffel function attains small values.

(ii) For domains D with bounded measure, we may choose $\beta(\boldsymbol{x}) = \varrho_T(D)$, as it was done in [10, 11, 29]. As all weights $\omega_i = \varrho_T(D)$, S_m becomes the plain least squares method. In this case, $\|\beta\|_\infty$ is minimal and noise is amplified the least. But this choice spoils the assumption on the choice of m when the Christoffel function $N(V_m, \boldsymbol{x})$ attains big values. This effect is controllable, for instance, when working with a bounded orthonormal system (BOS) ($\|\eta_k\|_\infty \leq B$ for some $B > 0$ and all k). Then

$$N(V_m) \leq \sum_{k=0}^{m-1} \|\eta_k\|_\infty^2 \leq mB^2$$

and the assumption on the size of V_m can be replaced by

$$10\|\beta(\cdot)N(V_m,\cdot)\|_\infty(\log(m)+t) \leq 10\varrho_T(D)Bm(\log(m)+t) \leq n\,.$$

An interesting example, where these effects occur, is the approximation of functions on the unit interval $D = [0,1]$ from samples given in uniformly random points.

- When using algebraic polynomials and the Lebesgue error measure $\mathrm{d}\varrho_T = \mathrm{d}x$ we have to choose $\beta \equiv 1$ to obtain uniform random points. Orthogonalizing algebraic polynomials with respect to the Lebesgue measure, we obtain $\eta_k = P_k/\|P_k\|_{L_2((0,1),\mathrm{d}x)}$ Legendre polynomials for our approximation space V_m. Since $\|P_k\|^2_{L_2((0,1),\mathrm{d}x)} = 2k+1$ and $P_k(0) = 1$, we have

$$N(V_m, 0) = \sum_{k=0}^{m-1} \frac{|P_k(0)|^2}{\|P_k\|^2_{L_2}} = \sum_{k=0}^{m-1} (2k+1) = m^2\,. \tag{1.1}$$

  Thus, this case falls into category (i) from above and spoils our choice of $m \leq \sqrt{n}$, i.e., quadratic oversampling as in [31].

- When using algebraic polynomials and the Chebyshev error measure $\mathrm{d}\varrho_T = (1-(2x-1)^2)^{-1/2}\,\mathrm{d}x$ we have to choose $\beta(x) = \frac{\pi}{4}(1-(2x-1)^2)^{-1/2}$ to obtain uniform random points. Orthogonalizing algebraic polynomials with respect to the Chebyshev measure, we obtain Chebyshev polynomials $\eta_k(x) = T_k(x) = \cos(k \arccos(2x-1))$ for our approximation space V_m. These are a BOS, but the distribution β spoils both the assumption on m and the error bound, since β diverges at the border (this effect can be circumvented using a padding technique at the border as we discuss in Section 4).

- In Section 4.1 we construct orthogonal functions with respect to the Sobolev space inner product

$$\langle f, g\rangle_{H^s(0,1)} = \langle f, g\rangle_{L_2((0,1),\mathrm{d}x)} + \langle f^{(s)}, g^{(s)}\rangle_{L_2((0,1),\mathrm{d}x)}\,,$$

  for $s = 1$ and $s = 2$, which are orthogonal with respect to the $L_2((0,1),\mathrm{d}x)$ inner product as well. For $s = 1$, these functions were originally introduced in [25] and for $s = 2$ in [3, Section 3], where also higher orders can be found.

  We show that they form a BOS and, by (ii) above, this basis is then suitable for approximation in uniform random points on $D = [0,1]$ using plain least squares and only logarithmic oversampling. The H^2 basis is prone to numerical errors. To overcome this, we propose a numerically stable approximation and proof its accuracy.

As for the structure of this paper, we start with some tools from probability theory in Section 2. In Section 3 we show error bounds for the weighted least squares method. The construction of the H^1 and H^2 basis mentioned above are found in Section 4 along with a comparison to the Legendre and Chebyshev polynomials. To indicate the applicability of our error bounds and the proposed basis, we conduct numerical experiments in one and five dimensions.

2. Tools from probabilty theory

For the later analysis we need concentration inequalities starting with Bernstein's inequality, which is found in the standard literature, cf. [45, Theorem 6.12] or [17, Corollary 7.31].

Theorem 2.1 (Bernstein). *Let $\xi_1, \dots, \xi_n$ be independent real-valued mean-zero random variables satisfying $\mathbb{E}(\xi_i^2) \le \sigma^2$ and $\|\xi_i\|_\infty \le B$ for $i = 1, \dots, n$ and real numbers σ^2 and B. Then*

$$\frac{1}{n}\sum_{i=1}^{n} \xi_i \le \frac{2Bt}{3n} + \sqrt{\frac{2\sigma^2 t}{n}}$$

with probability exceeding $1 - \exp(-t)$.

Bernstein's inequality gives a concentration bound for the sum of independent random variables. We need similar bounds for quadratic forms in random vectors, which are known as Hanson-Wright inequalities. To formulate them, we need to introduce the spectral norm and the Frobenius norm of a matrix $\boldsymbol{A} \in \mathbb{C}^{m\times n}$

$$\|\boldsymbol{A}\|_{2\to 2} = \sqrt{\lambda_{\max}(\boldsymbol{A}^*\boldsymbol{A})} = \sigma_{\max}(\boldsymbol{A}) \quad \text{and} \quad \|\boldsymbol{A}\|_F = \sqrt{\sum_{k=1}^{m}\sum_{i=1}^{n} |a_{k,i}|^2}\,,$$

where $\lambda_{\max}$ and $\sigma_{\max}$ denote the largest eigenvalues and singular values, respectively. The following result is such an inequality with a Bernstein condition on the random variables and was shown in [4, Theorem 3].

Theorem 2.2 (Hanson-Wright). *Let $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)^\mathsf{T}$ be a vector of independent mean-zero random variables such that for all integers $p \ge 1$*

$$\mathbb{E}(|\xi_i|^{2p}) \le p! B^{2p-2}\sigma_i^2/2 \tag{2.1}$$

for real numbers $B \ge 0$, $\sigma_i \ge 0$, and all $i = 1, \dots, n$. Let further $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ and $m = \mathbb{E}(\boldsymbol{\xi}^*\boldsymbol{A}\boldsymbol{\xi})$. Then $\boldsymbol{D}_\sigma = \operatorname{diag}(\sigma_1, \dots, \sigma_n)$

$$\boldsymbol{\xi}^*\boldsymbol{A}\boldsymbol{\xi} - m \le 256B^2\|\boldsymbol{A}\boldsymbol{D}_\sigma\|_{2\to 2}t + 8\sqrt{3}B\|\boldsymbol{A}\boldsymbol{D}_\sigma\|_F\sqrt{t}$$

with probability exceeding $1 - \exp(-t)$.

The following is a special case of the above Hanson-Wright inequality for Hermitian positive semi-definite matrices and random variables with known variance $\mathbb{E}(|\xi_i|^2)$ and a uniform bound $\|\xi_i\|_\infty$.

Corollary 2.3. *Let $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)^\mathsf{T}$ be a vector of independent complex-valued mean-zero random variables satisfying $\mathbb{E}(|\xi_i|^2) \leq \sigma^2$ and $\|\xi_i\|_\infty \leq B$ for $i = 1, \dots, n$. Then for all $\boldsymbol{A} \in \mathbb{C}^{m\times n}$*

$$\|\boldsymbol{A}\boldsymbol{\xi}\|_2^2 \leq 128B^2\|\boldsymbol{A}\|_{2\to 2}^2 t + (8\sqrt{3}B\sqrt{t\sigma^2} + \sigma^2)\|\boldsymbol{A}\|_F^2$$

with probability exceeding $1 - \exp(-t)$.

Proof. Since $\|\boldsymbol{A}\boldsymbol{\xi}\|_2^2 = \boldsymbol{\xi}^*\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{\xi}$ is a quadratic form we want to apply Theorem 2.2 on $\boldsymbol{A}^*\boldsymbol{A}$. For that we check the moment condition (2.1) on $\xi_1^2, \dots, \xi_n^2$. For $p = 1$ it is fulfilled for constants $B/\sqrt{2}$ and $(\sqrt{2}\sigma_i)^2$. For $p \geq 2$, we have $p! \geq 2^{p-1}$ and obtain

$$\begin{aligned}\mathbb{E}(|a_i\xi_i|^{2p}) &\leq \|\xi_i\|_\infty^{2p-2}\mathbb{E}(|\xi_i|^2)\\ &\leq (B)^{2p-2}\sigma^2\\ &\leq p!\Big(\frac{B}{\sqrt{2}}\Big)^{2p-2}\frac{(\sqrt{2}\sigma)^2}{2}\,.\end{aligned}$$

Therefore, Theorem 2.2 is applicable.
It is left to estimate the expected value. Since $\xi_1, \dots, \xi_n$ are independent and have bounded variance, we obtain

$$\begin{aligned}\mathbb{E}\Big(\|\boldsymbol{A}\boldsymbol{\xi}\|_2^2\Big) &= \sum_{k=1}^m\sum_{i=1}^n\sum_{j=1}^n a_{i,k}\overline{a_{j,k}}\,\mathbb{E}(\xi_i\overline{\xi_j})\\ &= \sum_{k=1}^m\Big(\sum_{i=1}^n\sum_{j\neq i} a_{i,k}\overline{a_{j,k}}\,\mathbb{E}(\xi_i\overline{\xi_j})\Big) + \sum_{i=1}^n|a_{i,k}|^2\mathbb{E}(|\xi_i|^2)\\ &\leq \sigma^2\|\boldsymbol{A}\|_F^2\,.\end{aligned}$$

■

The following tool is a concentration bound on the maximal singular values of random matrices which was shown in [51, Theorem 1.1].

Lemma 2.4 (Matrix Chernoff)**.** *For a finite sequence $\boldsymbol{A}_1, \dots, \boldsymbol{A}_n \in \mathbb{C}^{m\times m}$ of independent, Hermitian, positive semi-definite random matrices satisfying $\lambda_{\max}(\boldsymbol{A}_i) \leq R$ almost surely it holds*

$$\begin{aligned}\mathbb{P}\Big(\lambda_{\min}\Big(\sum_{i=1}^n\boldsymbol{A}_i\Big) \leq (1-t)\mu_{\min}\Big) &\leq m\exp\Big(-\frac{\mu_{\min}}{R}(t + (1-t)\log(1-t))\Big)\\ &\leq m\exp\Big(-\frac{\mu_{\min}t^2}{2R}\Big)\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}\Big(\lambda_{\max}\Big(\sum_{i=1}^n\boldsymbol{A}_i\Big) \geq (1+t)\mu_{\max}\Big) &\leq m\exp\Big(-\frac{\mu_{\max}}{R}(-t + (1+t)\log(1+t))\Big)\\ &\leq m\exp\Big(-\frac{\mu_{\max}t^2}{3R}\Big)\end{aligned}$$

for $t \in [0, 1]$ where $\mu_{\min} \coloneqq \lambda_{\min}(\sum_{i=1}^n \mathbb{E}(\boldsymbol{A}_i))$ and $\mu_{\max} \coloneqq \lambda_{\max}(\sum_{i=1}^n \mathbb{E}(\boldsymbol{A}_i))$.

Proof. The first estimates are provided by [51, Theorem 1.1]. Based on the Taylor expansion

$$(1+t)\log(1+t) = t + \sum_{k=2}^{\infty} \frac{(-1)^k}{k(k-1)} t^k \,,$$

which holds true for $t \in [-1,1]$, we further derive the inequalities

$$t + (1-t)\log(1-t) = \sum_{k=2}^{\infty} \frac{1}{k(k-1)} t^k \geq \frac{t^2}{2}$$

and

$$-t + (1+t)\log(1+t) = \sum_{k=2}^{\infty} \frac{(-1)^k}{k(k-1)} t^k \geq \frac{t^2}{2} - \frac{t^3}{6} \geq \frac{t^2}{3}$$

for the range $t \in [0,1]$. ∎

3. **Error bounds for least squares**

In this section we develop L_2- and L_∞-error bounds for the least squares method. To this end we introduce some notation and the method itself. Let $\eta_0, \dots, \eta_{m-1} \colon D \to \mathbb{C}$ be an L_2-orthonormal basis of V_m,

$$N(V_m, \boldsymbol{x}) = \sum_{k=0}^{m-1} |\eta_k(\boldsymbol{x})|^2 \quad \text{and} \quad N(V_m) = \sup_{\boldsymbol{x} \in D} N(V_m, \boldsymbol{x})$$

be the *Christoffel function* and its supremum. For our approximation method S_m we use the *weighted least squares approximation* depending on $\eta_0, \dots, \eta_{m-1}$ and $\boldsymbol{x}^1, \dots, \boldsymbol{x}^n$:

$$(S_m \boldsymbol{y})(\boldsymbol{x}) = \sum_{k=0}^{m-1} \hat{g}_k \eta_k(\boldsymbol{x}) \quad \text{with} \quad \hat{\boldsymbol{g}} = \operatorname*{arg\,min}_{\hat{\boldsymbol{a}} \in \mathbb{C}^m} \|\boldsymbol{L}\hat{\boldsymbol{a}} - \boldsymbol{y}\|_{\boldsymbol{W}}^2 \,,$$

$$\boldsymbol{L} = \begin{bmatrix} \eta_0(\boldsymbol{x}^1) & \dots & \eta_{m-1}(\boldsymbol{x}^1) \\ \vdots & \ddots & \vdots \\ \eta_0(\boldsymbol{x}^n) & \dots & \eta_{m-1}(\boldsymbol{x}^n) \end{bmatrix} \in \mathbb{C}^{n \times m}, \text{ and } \quad \boldsymbol{W} = \begin{bmatrix} \omega_1 & & \\ & \ddots & \\ & & \omega_n \end{bmatrix} \in [0,\infty)^{n \times n} \tag{3.1}$$

where $\|\boldsymbol{L}\hat{\boldsymbol{a}} - \boldsymbol{y}\|_{\boldsymbol{W}}^2 = (\boldsymbol{L}\hat{\boldsymbol{a}} - \boldsymbol{y})^* \boldsymbol{W} (\boldsymbol{L}\hat{\boldsymbol{a}} - \boldsymbol{y})$. If all weights are equal we speak of *plain least squares approximation.*

The coefficients $\hat{\boldsymbol{g}}$ of the approximation $S_m \boldsymbol{y}$ are found by solving the normal equation

$$\hat{\boldsymbol{g}} = (\boldsymbol{L}^* \boldsymbol{W} \boldsymbol{L})^{-1} \boldsymbol{L}^* \boldsymbol{W} \boldsymbol{y} \,.$$

Doing this by the means of an iterative solver, the stability and the iteration count for a desired precision are determined by the spectral properties of the above matrix, cf. [19, Theorem 3.1.1]. However, these are fully determined by the spectral properties of $\boldsymbol{W}^{1/2}\boldsymbol{L}$, since for a singular value decomposition $\boldsymbol{W}^{1/2}\boldsymbol{L} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$, we obtain

$$(\boldsymbol{L}^* \boldsymbol{W} \boldsymbol{L})^{-1} \boldsymbol{L}^* \boldsymbol{W}^{1/2} = \boldsymbol{V} (\boldsymbol{\Sigma}^* \boldsymbol{\Sigma})^{-1} \boldsymbol{\Sigma} \boldsymbol{U}^* \,. \tag{3.2}$$

For $f = \sum_{k=0}^{m-1} \hat{f}_k \eta_k \in V_m$, the singular values of $\boldsymbol{W}^{1/2}\boldsymbol{L}$ relate the coefficients $\hat{\boldsymbol{f}} = (\hat{f}_0, \dots, \hat{f}_{m-1})^\mathsf{T}$ with the samples $\boldsymbol{f} = (f(\boldsymbol{x}^1), \dots, f(\boldsymbol{x}^n))^\mathsf{T} = \boldsymbol{L}\hat{\boldsymbol{f}}$. Such connection is known as L_2-Marcinkiewicz-Zygmund inequality for V_m. It was established in [11, Thm. 2.1] that random points also fulfill this, which is central in all theorems presented. This makes it applicable in a very general setting, cf. [33, Theorem 2.3], [32, Theorem 5.1], [15, Lemma 2.1], or [8, Theorem 2.1].

Lemma 3.1. *Let $t \geq 0$, $n \in \mathbb{N}$, $\boldsymbol{x}^1, \dots, \boldsymbol{x}^n$ be points drawn according to a probability measure $\mathrm{d}\varrho_S = 1/\beta \, \mathrm{d}\varrho_T$. Let further, V_m be an m-dimensional function space with an orthonormal basis $\eta_0, \dots, \eta_{m-1}$ in L_2 with m satisfying*

$$10\|\beta(\cdot)N(V_m,\cdot)\|_\infty(\log(m)+t) \leq n$$

and $\boldsymbol{L}, \boldsymbol{W}$ be as in (3.1) with $\omega_i = \beta(\boldsymbol{x}^i)$. Then

$$\frac{n}{2}\|\hat{\boldsymbol{g}}\|_2^2 \leq \|\boldsymbol{W}^{1/2}\boldsymbol{L}\hat{\boldsymbol{g}}\|_2^2 \leq \frac{3n}{2}\|\hat{\boldsymbol{g}}\|_2^2 \quad \text{for all} \quad \hat{\boldsymbol{g}} \in \mathbb{C}^m,$$

where each inequality holds with probability exceeding $1-\exp(-t)$, repectively.

The proof ideas go back to [10, Thm. 1] and [11, Thm. 2.1] but for the sake of readability we state it here as well.

Proof. The result is a direct consequence of Tropp's result in Lemma 2.4. For a randomly chosen point $\boldsymbol{x}^i$ we define the random rank-one matrix $\boldsymbol{A}_i = \frac{1}{n}\beta(\boldsymbol{x}^i)(\boldsymbol{y}^i \otimes \boldsymbol{y}^i)$ with $\boldsymbol{y}^i = (\eta_0(\boldsymbol{x}^i), \dots, \eta_{m-1}(\boldsymbol{x}^i))^\mathsf{T}$. By construction, it holds

$$\sum_{i=1}^n \boldsymbol{A}_i = \boldsymbol{L}^*\boldsymbol{W}\boldsymbol{L}$$

and by the orthogonality of η_k

$$\Big(\mathbb{E}(\boldsymbol{A}_i)\Big)_{k,l} = \frac{1}{n}\int_D \eta_k(\boldsymbol{x})\overline{\eta_l(\boldsymbol{x})}\beta(\boldsymbol{x})\beta^{-1}(\boldsymbol{x}) \, d\varrho_T(\boldsymbol{x}) = \frac{\delta_{k,l}}{n},$$

which gives $\mathbb{E}\Big(\sum_{i=1}^n \boldsymbol{A}_i\Big) = \mathrm{Id}_{m\times m}$ and, therefore, $\mu_{\max} = \mu_{\min} = 1$. Further, we have

$$\lambda_{\max}\Big(\frac{1}{n}\beta(\boldsymbol{x}^i)(\boldsymbol{y}^i \otimes \boldsymbol{y}^i)\Big) = \frac{1}{n}\beta(\boldsymbol{x}^i)\|\boldsymbol{y}^i\|_2^2 \leq \frac{1}{n}\|\beta(\cdot)N(V_m,\cdot)\|_\infty .$$

Lemma 2.4 with $t = 1/2$ then gives the lower bound

$$\mathbb{P}\Big(\lambda_{\min}\Big(\frac{1}{n}\boldsymbol{L}^*\boldsymbol{W}\boldsymbol{L}\Big) \leq \frac{1}{2}\Big) \leq m\exp\Big(-\frac{n}{10}\|\beta(\cdot)N(V_m,\cdot)\|_\infty^{-1}\Big),$$

which is smaller than $\exp(-t)$ by the assumption on m.
The bound for the largest eigenvalue works analogue. ∎

We now formulate a bound on the L_2-error of the weighted least squares method for exact function values. This result is heavily based on [29, Theorem 3.20] which extends to a more general setting.

Theorem 3.2 (L_2-error bound without noise)**.** *Let $f\colon D \to \mathbb{C}$, $\boldsymbol{x}^1, \dots, \boldsymbol{x}^n$, $n \in \mathbb{N}$ be points drawn according to a probability measure $\mathrm{d}\varrho_S = 1/\beta \, \mathrm{d}\varrho_T$ and $\boldsymbol{y} = (f(\boldsymbol{x}^1), \dots, f(\boldsymbol{x}^n))^\mathsf{T}$ exact function values. Let further, $t \geq 0$, V_m be an m-dimensional function space with an orthonormal basis $\eta_0, \dots, \eta_{m-1}$ satisfying*

$$10\|\beta(\cdot)N(V_m,\cdot)\|_\infty(\log(m)+t) \leq n .$$

Then, for S_m the weighted least squares method defined in (3.1) with $\omega_i = \beta(\boldsymbol{x}^i)$, we have with probability exceeding $1-2\exp(-t)$:

$$\begin{aligned}\|f - S_m\boldsymbol{y}\|_{L_2}^2 &\leq 8\Big(e(f,V_m,L_2)_{L_2} + \sqrt{\frac{t}{n}}e(f,V_m,L_2)_{L_\infty}\Big)^2 \\ &\leq 8\Big(1+\sqrt{\frac{N(V_m)}{\|\beta(\cdot)N(V_m,\cdot)\|_\infty}}\Big)^2 e(f,V_m,L_2)_{L_2}^2 .\end{aligned}$$

Proof. For abbreviation, we use $e_2 = e(f, V_m, L_2)_{L_2}$ and $e_\infty = e(f, V_m, L_2)_{L_\infty}$. Further, we define the event

$$A \coloneqq \Big\{ \boldsymbol{x}^1, \dots, \boldsymbol{x}^n \in D : \frac{n}{2} \leq \|\boldsymbol{W}^{1/2}\boldsymbol{L}\|_{2\to 2}^2 \Big\} \tag{3.3}$$

which has probability $\mathbb{P}(A) \geq 1 - \exp(-t)$ by Lemma 3.1 and the assumption on V_m. We split the approximation error

$$\|f - S_m \boldsymbol{f}\|_{L_2}^2 = e_2^2 + \|P(f, V_m, L_2) - S_m \boldsymbol{f}\|_{L_2}^2 .$$

Due to the invariance of S_m to functions in V_m, we pull it in front and use compatibility of the operator norm to obtain

$$\|f - S_m \boldsymbol{f}\|_{L_2}^2 \leq e_2^2 + \|S_m\|_{2\to 2}^2 \sum_{i=1}^n \beta(\boldsymbol{x}^i) |(f - P(f, V_m, L_2))(\boldsymbol{x}^i)|^2 .$$

By (3.2) and the event (3.3), we have $\|S_m\|_{2\to 2}^2 = \|\boldsymbol{W}^{1/2}\boldsymbol{L}\|_{2\to 2}^{-1} \leq 2/n$. Thus,

$$\|f - S_m \boldsymbol{f}\|_{L_2}^2 \leq 3e_2^2 + \frac{2}{n} \sum_{i=1}^n \Big| |\omega_i (f - P(f, V_m, L_2))(\boldsymbol{x}^i)|^2 - e_2^2 \Big| .$$

It remains to estimate the latter summand. We define

$$\xi_i = \beta(\boldsymbol{x}^i) \Big| (f - P_m f)(\boldsymbol{x}^i) \Big|^2 - e_2^2,$$

which is mean-zero since we sample with respect to the distribution ϱ_S. Further, we have

$$\mathbb{E}(\xi_i^2) = \mathbb{E}\Big((\beta(\boldsymbol{x}^1))^2 |(f - P_m f)(\boldsymbol{x}^i)|^4 \Big) - e_2^4 \leq \|f - P_m f\|_{L_\infty}^2 e_2^2 - e_2^4 \leq e_2^2 (e_2 + e_\infty)^2 ,$$

and

$$\|\xi_i\|_\infty \leq \sup_{x \in D} \Big| \beta(\boldsymbol{x}) |(f - P_m f)(\boldsymbol{x})|^2 - e_2^2 \Big| \leq e_\infty^2 + e_2^2 .$$

Thus, the conditions in order to apply Bernstein are fulfilled:

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \xi_i &\leq \frac{2t}{3n} \Big(e_2^2 + e_\infty^2 \Big) + \sqrt{\frac{2t}{n}} \Big(e_\infty e_2 + e_2^2 \Big) \\ &\leq \Big(\frac{2}{3} + \sqrt{2} \Big) e_2^2 + \sqrt{\frac{2t}{n}} e_\infty e_2 + \frac{2t}{3n} e_\infty^2 \end{aligned} \tag{3.4}$$

with probability $1 - \exp(-t)$, where $t \leq n$ was used in the last inequality. Thus,

$$\begin{aligned} \|f - S_m \boldsymbol{f}\|_{L_2}^2 &\leq \Big(\frac{13}{3} + 2\sqrt{2} \Big) e_2^2 + \sqrt{\frac{8t}{n}} e_\infty e_2 + \frac{4t}{3n} e_\infty^2 \\ &\leq \Big(\frac{13}{3} + 2\sqrt{2} \Big) \Big(e_2 + \sqrt{\frac{t}{n}} e_\infty \Big)^2 . \end{aligned}$$

By union bound we obtain the overall probability exceeding the sum of the probabilities of events given by (3.3) and (3.4).

The second bound is attained as follows: For any $g = \sum_{k=1}^m \langle g, \eta_k \rangle_{L_2} \eta_k \in V_m$ the Hölder-inequality gives an estimate on the L_∞-norm in terms of the L_2-norm:

$$\|g\|_{L_\infty} = \Big\| \sum_{k=1}^m \langle g, \eta_k \rangle_{L_2} \eta_k \Big\|_{L_\infty} \leq \Big\| \sqrt{\sum_{k=1}^m |\langle g, \eta_k \rangle_{L_2}|^2} \sqrt{\sum_{k=1}^m |\eta_k|^2} \Big\|_{L_\infty} = \sqrt{N(V_m)} \|g\|_{L_2} . \tag{3.5}$$

Using the assumption on V_m, we have

$$\begin{aligned}\sqrt{\frac{t}{n}}e(f,V_m,L_2)_{L_\infty} &\le \sqrt{\frac{tN(V_m)}{n}}e(f,V_m,L_2)_{L_2}\\ &\le \sqrt{\frac{t}{10(\log(m)+t)}\frac{N(V_m)}{\|\beta(\cdot)N(V_m,\cdot)\|_\infty}}e(f,V_m,L_2)_{L_2}\\ &\le \sqrt{\frac{N(V_m)}{\|\beta(\cdot)N(V_m,\cdot)\|_\infty}}e(f,V_m,L_2)_{L_2}\,.\end{aligned}$$

■

Provided $N(V_m)/\|\beta(\cdot)N(V_m,\cdot)\|_\infty$ is finite, Theorem 3.2 says that the least squares approximation from a finite-dimensional function space V_m and given the oversampling condition has the same error as the L_2-projection up to a multiplicative constant with high probability. This improves on [11, Theorem 2.1] where the same bound was shown in expectation or bounded by the L_∞-error with high probability.

Next, we prove the central theorem which includes the noisy case.

Proof. [Proof of Theorem 1.1] We split the approximation error

$$\|f-S_m\boldsymbol{y}\|_{L_2}^2 \le e(f,V_m,L_2)_{L_2} + 2\|f-S_m\boldsymbol{f}\|_{L_2}^2 + 2\|S_m\boldsymbol{\varepsilon}\|_{L_2}^2$$

and bound the first two summands as in the proof of Theorem 3.2 with the events given by (3.3) and (3.4). Note, the constant changes from $13/3+2\sqrt{2}$ to $23/3+4\sqrt{2}\le 14$. Now, we focus on the third summand. Applying Corollary 2.3 gives

$$\begin{aligned}&\|S_m\boldsymbol{\varepsilon}\|_{L_2}^2 = \|(\boldsymbol{L}^*\boldsymbol{W}\boldsymbol{L})^{-1}\boldsymbol{L}^*\boldsymbol{W}\boldsymbol{\varepsilon}\|_2^2\\ &\le 128\|\boldsymbol{\varepsilon}\|_\infty^2\|(\boldsymbol{L}^*\boldsymbol{W}\boldsymbol{L})^{-1}\boldsymbol{L}^*\boldsymbol{W}\|_{2\to2}^2 t + (8\sqrt{3}\|\boldsymbol{\varepsilon}\|_\infty\sqrt{t\sigma^2}+\sigma^2)\|(\boldsymbol{L}^*\boldsymbol{W}\boldsymbol{L})^{-1}\boldsymbol{L}^*\boldsymbol{W}\|_F^2 \qquad (3.6)\end{aligned}$$

with probability $1-\exp(-t)$. Since $\boldsymbol{L}^*\boldsymbol{W}\boldsymbol{L}\in\mathbb{C}^{m\times m}$, the matrix $(\boldsymbol{L}^*\boldsymbol{W}\boldsymbol{L})^{-1}\boldsymbol{L}\boldsymbol{W}^{1/2}$ has rank at most m and, thus, we use $\|\boldsymbol{A}\|_F^2\le\operatorname{rank}(\boldsymbol{A})\|\boldsymbol{A}\|_{2\to2}^2$ to obtain

$$\|(\boldsymbol{L}^*\boldsymbol{W}\boldsymbol{L})^{-1}\boldsymbol{L}^*\boldsymbol{W}\|_F^2 \le \|\beta\|_\infty m\|(\boldsymbol{L}^*\boldsymbol{W}\boldsymbol{L})^{-1}\boldsymbol{L}^*\boldsymbol{W}^{1/2}\|_{2\to2}^2 \le \|\beta\|_\infty\frac{2m}{n}$$

where the last inequality follows from (3.2) and event (3.3). Therefore,

$$\|S_m\boldsymbol{\varepsilon}\|_{L_2}^2 \le \|\beta\|_\infty\Big(128\|\boldsymbol{\varepsilon}\|_\infty^2\frac{2}{n}t + (8\sqrt{3}\|\boldsymbol{\varepsilon}\|_\infty\sqrt{t\sigma^2}+\sigma^2)\frac{2m}{n}\Big)\,.$$

By union bound we obtain the overall probability exceeding the sum of the probabilities of the events given by (3.3), (3.4), and (3.6). ■

Next, we prove Theorem 1.2 bounding the approximation error of least squares in the L_∞-norm.

Proof. [Proof of Theorem 1.2] For abbreviation, we use $e_2 = e(f,V_m,L_\infty)_{L_2}$ and $e_\infty = e(f,V_m,L_\infty)_{L_\infty}$. Using (3.5) we reduce the L_∞-case to the L_2-case which we already covered. We split the approximation error

$$\begin{aligned}\|f-S_m\boldsymbol{y}\|_{L_\infty} &\le \|f-P(f,V_m,L_\infty)\|_{L_\infty} + \|P(f,V_m,L_\infty)-S_m\boldsymbol{f}\|_{L_\infty} + \|S_m\boldsymbol{\varepsilon}\|_{L_\infty}\\ &\le e_\infty + \sqrt{N(V_m)}\|P(f,V_m,L_\infty)-S_m\boldsymbol{f}\|_{L_2} + \sqrt{N(V_m)}\|S_m\boldsymbol{\varepsilon}\|_{L_2}\,.\end{aligned}$$

Analogously to (3.4) we obtain

$$\begin{aligned}\|P(f,V_m,L_\infty)-S_m\boldsymbol{f}\|_{L_2}^2 &\le \Big(\frac{2}{3}+\sqrt{2}\Big)e_\infty^2+\sqrt{\frac{2t}{n}}e_\infty e_2+\frac{2t}{3n}e_2^2\\ &\le \Big(\frac{2}{3}+\sqrt{2}\Big)\Big(e_\infty+\sqrt{\frac{t}{n}}e_2\Big)^2,\end{aligned}$$

where the last inequality follows from $t \le n$. Thus,

$$\|f-S_m\boldsymbol{y}\|_{L_\infty} \le \Big(1+\sqrt{\frac{4+6\sqrt{2}}{3}N(V_m)}\Big)\Big(e_\infty+\sqrt{\frac{t}{n}}e_2\Big)+\sqrt{N(V_m)}\|S_m\boldsymbol{\varepsilon}\|_{L_2}\,.$$

Using the same bound as in Theorem 1.1 for $\|S_m\boldsymbol{\varepsilon}\|_{L_2}$ we obtain the assertion. ∎

4. **Application on the unit cube**

In this section we are interested in approximating functions on the d-dimensional unit cube $D = [0,1]^d$ when sample points are drawn uniformly, i.e., with respect to the Lebesgue measure $\mathrm{d}\boldsymbol{x}$. The deterministic equivalent to uniform sampling are equispaced points. When using these for polynomial interpolation, Runge already knew in 1901, that higher degree polynomials lead to oscillatory behaviour towards the border which spoil the approximation error. Even though, we do not interpolate, we will observe similar behaviour using Legendre and Chebyshev polynomials. We propose alternative bases, which are stable for large $m = \dim(V_m)$ as well.

Throughout this section we have $L_2 = L_2((0,1)^d, \mathrm{d}\boldsymbol{x})$ unless stated otherwise.

We consider function spaces to know about the decay of the coefficients. Note, that for individual functions they may decay faster in contrast to the worst-case setting. Literature for the worst-case setting can be found in the papers [26] for random points with logarithmic oversampling, [33, 28] for subsampled points with linear oversampling and a logarithmic gap in the error bound (this was made constructive in [8]), and [16] for subsampled points with linear oversampling and sharp bounds.

4.1. **Sobolev spaces on the unit interval**

Let $d = 1$, $D = [0,1]$ be the unit interval equipped with the Lebesgue measure $\mathrm{d}x$. In order to get hold on appropriate finite-dimensional function spaces V_m for approximation, we have a look at Sobolev spaces $H^s = H^s(0,1)$ with integer smoothness $s \ge 0$. The inner product of these Hilbert spaces is given by

$$\langle f,g\rangle_{H^s} = \langle f,g\rangle_{L_2} + \langle f^{(s)},g^{(s)}\rangle_{L_2}\,.$$

Since $\|f\|_{L_2}^2 \le \|f\|_{H^s}^2 = \langle f,f\rangle_{H^s}$, the embedding operator $\mathrm{Id}\colon H^s \hookrightarrow L_2$ is compact. Thus, $W = \mathrm{Id}^* \circ \mathrm{Id}\colon H^s \to H^s$ is compact and self-adjoint. Applying the spectral theorem gives for $f \in H^s$

$$W(f) = \sum_{k=0}^{\infty} \sigma_k \langle f, e_k\rangle_{H^s} e_k$$

where $(\sigma_k)_{k=0}^\infty$ is the non-increasing rearrangement of the singular values of W and $(e_k)_{k=0}^\infty \subset H^s$ the corresponding system of eigenfunctions forming an orthonormal basis in H^s. Since

$$\langle e_k,e_l\rangle_{L_2} = \langle \mathrm{Id}(e_k),\mathrm{Id}(e_l)\rangle_{L_2} = \langle W(e_k),e_l\rangle_{H^s} = \sigma_k^2\langle e_k,e_l\rangle_{H^s} = \sigma_k^2\delta_{k,l}\,,$$

the functions $\eta_k = \sigma_k^{-1} e_k$ form an orthonormal system in L_2. Setting $V_m = \operatorname{span}\{\eta_k\}_{k=0}^{m-1}$, we obtain for $H^s \ni f = \sum_{k=0}^{\infty} \langle f, e_k\rangle_{H^s} e_k$

$$e(f, V_m, L_2)_{L_2}^2 = \Big\| \sum_{k=m}^{\infty} \langle f, e_k\rangle_{H^s} e_k \Big\|_{L_2}^2 = \sum_{k=m}^{\infty} \Big| \langle f, e_k\rangle_{H^s} \sigma_k \Big|^2 \leq \|f\|_{H^s}^2 \sigma_m^2 \,. \tag{4.1}$$

Thus, the eigenfunctions corresponding to the largest singular values are a canonical candidate for the approximation space V_m. To put this into concrete terms, in the next two theorems, we give the singular values and eigenfunctions for H^1 and H^2.

Theorem 4.1. *The operator $W = \mathrm{Id}^* \circ \mathrm{Id} \colon H^1 \to H^1$ has singular values $\sigma_k^2 = \frac{1}{1+\pi^2 k^2}$ with corresponding L_2-normalized eigenfunctions*

$$\eta_k(x) = \begin{cases} 1 & \text{for } k = 0 \\ \sqrt{2}\cos(\pi k x) & \text{for } k \geq 1 \,. \end{cases}$$

Proof. For σ a singular value of W with corresponding eigenfunction $\eta \in H^1$ and $\varphi \in H^1$ a test function, we have

$$\langle \eta, \varphi\rangle_{L_2} = \langle \mathrm{Id}(\eta), \mathrm{Id}(\varphi)\rangle_{L_2} = \langle W(\eta), \varphi\rangle_{H^1} = \sigma^2 \langle \eta, \varphi\rangle_{H^1} = \sigma^2 \Big(\langle \eta, \varphi\rangle_{L_2} + \langle \eta', \varphi'\rangle_{L_2} \Big) \,.$$

Partial differentiation yields $\langle \eta', \varphi'\rangle_{L_2} = \eta'(1)\varphi(1) - \eta'(0)\varphi(0) - \langle \eta'', \varphi\rangle_{L_2}$. Thus

$$\Big\langle \frac{1-\sigma^2}{\sigma^2}\eta + \eta'', \varphi \Big\rangle_{L_2} = \eta'(1)\varphi(1) - \eta'(0)\varphi(0) \,.$$

Since this has to hold for all test functions $\varphi \in H^1$, we obtain the differential equation

$$\frac{1-\sigma^2}{\sigma^2}\eta = -\eta'' \quad \text{with} \quad \eta(0)' = \eta(1)' = 0 \,.$$

The proposed functions are exactly the ones fulfilling this differential equation. ∎

To our knowledge, the H^1 basis above was originally introduced in [25] and was already considered in [53, Lemma 4.1] with the same proof technique, in [23] as a modified Fourier expansion. It is further used in [46] as a basis for multivariate approximation in the context of samples along tent-transformed rank-1 lattices, and in [1, 2, 13, 12, 27]. The following H^2 basis was already posed in [3, Section 3], where higher-order Sobolev-spaces are found as well. The proof of the following theorem is found in Appendix A.

Theorem 4.2. *The operator $W = \mathrm{Id}^* \circ \mathrm{Id} \colon H^2 \to H^2$ has singular values $\sigma_0^2 = 1$ with corresponding L_2-normalized eigenfunctions*

$$\eta_0(x) = 1 \quad \textit{and} \quad \eta_1(x) = 2\sqrt{3}x - \sqrt{3}$$

and for $k \geq 2$, $\sigma_k^2 = \frac{1}{1+t_k^4}$ with $t_k > 0$ the solutions of $\cosh(t_k)\cos(t_k) = 1$ ($t_k \approx \frac{2k-1}{2}\pi$, cf. Lemma A.1) and

$$\eta_k(x) = \cosh(t_k x) + \cos(t_k x) - \frac{\cosh(t_k) - \cos(t_k)}{\sinh(t_k) - \sin(t_k)} (\sinh(t_k x) + \sin(t_k x)) \,.$$

Further, it holds

$$\|\eta_k\|_\infty \leq \begin{cases} 1 & \text{for } k = 0 \\ \sqrt{3} & \text{for } k = 1 \\ \sqrt{6} & \text{for } k \geq 2 \,. \end{cases}$$

The singular values σ_k for H^2 decay quadratic in contrast to linearly for H^1. Thus, approximating a twice differentiable function, $m = \dim(V_m)$ can be chosen smaller when using the H^2 basis whilst achieving the same truncation error $e(f, V_m, L_2)_{L_2}$. Furthermore, as noise enters with the factor m/n, cf. Theorem 1.1, this helps prevent overfitting as well and leads to a smaller approximation error.

However, as cosh and sinh both grow exponentially, the representation of the H^2 basis in Theorem 4.2 is prone to cancellations and, therefore, numerical unstable. In the next theorem we pose an approximation which is numerically stable.

Theorem 4.3. *For $0 < t_2 < t_3 < \dots$ fulfilling $\cosh(t_k)\cos(t_k) = 1$, let η_k be as in Theorem 4.2. Further, for $n \geq 2$, let $\tilde{t}_k = \pi(2k-1)/2$ and*

$$\begin{aligned}\tilde{\eta}_k(x) &= \sqrt{2}\cos\left(\tilde{t}_k x + \pi/4\right) \\ &\quad + \mathbb{1}_{[0,1/2]}(x)\exp\left(-\tilde{t}_k x\right) + \mathbb{1}_{[1/2,1]}(x)(-1)^k \exp\left(-\tilde{t}_k(1-x)\right).\end{aligned}$$

Then $|\eta_k(x) - \tilde{\eta}_k(x)| \leq \varepsilon$ for $k \geq \frac{2}{\pi}\log(16/\varepsilon) + 1$. In particular, the approximation $\tilde{\eta}_k$ is exact up to machine precision $\varepsilon = 10^{-16}$ for $k \geq 27$.

For the proof see Appendix A. For the numerical experiments we use the exact representation from Theorem 4.2 for $m < 10$ and the approximation from Theorem 4.3 for $m \geq 10$.

4.2. **Polynomial approximation on the unit interval**

Next, we examine how the H^1 and H^2 bases compares to polynomial approximation when points are distributed uniformly, i.e., $V_m = \Pi_m = \operatorname{span}\{1, x, \dots, x^{m-1}\}$ and $\varrho_S(x) = 1/\beta(x) d\varrho_T(x) = \mathrm{d}x$. Polynomial approximation results often assume $f \in X^s$ with

$$X^s \coloneqq \{f\colon [0,1] \to \mathbb{C} : f, \dots, f^{(s-1)} \text{ absolute continuous}, f^{(s)} \in BV([0,1])\},$$

where $BV([0,1])$ are all functions with bounded variation. This assumtion is stronger than assuming $f \in H^s$ as the following remark shows.

Remark 4.4 ($X^s \hookrightarrow H^{s+1/2-\varepsilon}$)**.** For a rigorous investigation of the relation of X^s and H^s, we need to define the Besov space $B^s_{p,q}$ for $p = 1$, $q = \infty$, and integer smoothness s

$$B^s_{1,\infty} \coloneqq \left\{ f \in L_1 : \sup_{h \neq 0} \frac{\|\Delta_h^2 f^{(s-1)}\|_{L_1}}{|h|} < \infty \right\}$$

with the finite difference $(\Delta_h f)(x) \coloneqq f(x+h) - f(x)$ and $\Delta_h^2 = \Delta_h \circ \Delta_h$, cf. [50, Section 1.2.5].

For $f \in X^s$ the derivative $f^{(s)}$ is of bounded variation. Thus, also the finite difference $\Delta_h^2 f$ is of bounded variation. In particular, $f^{(s)} \in L_1$ and, therefore, $f \in B^{s+1}_{1,\infty}$. By [50, (2.3.2/23)], we further have $B^{s+1}_{1,\infty} \hookrightarrow B^{s+1-\varepsilon}_{1,1}$ for any $\varepsilon > 0$. Thus,

$$X^s \hookrightarrow B^{s+1}_{1,\infty} \hookrightarrow B^{s+1-\varepsilon}_{1,1} \hookrightarrow H^{s+1/2-\varepsilon},$$

where the third embedding follows from the Sobolev inequality, cf. [49, (2.7.1/1)], and the Sobolev space for non-integer smoothness s consists of functions f such that $\langle f, \eta_k \rangle \leq Ck^{-s}$ for some constant $C < \infty$.

Assuming $f \in X^s$, we have a look into approximating with Legendre- and Chebyshev polynomials:

- The canonical choice for the target measure is $\mathrm{d}\varrho_T = \mathrm{d}x$ and $\beta \equiv 1$ such that $\varrho_S(x) = 1/\beta(x)\mathrm{d}\varrho_T(x) = \mathrm{d}x$. Orthogonalizing the first $m-1$ monomials with respect to $\mathrm{d}\varrho_T(x) = \mathrm{d}x$, we obtain the Legendre polynomials P_k.

For the error of the projection, assuming $f \in H^s$, the following was shown in [52, Theorem 3.5]:

$$e(f, V_m, L_2)_{L_2}^2 \leq \frac{2V}{\pi(s+1/2)(m-s)^{2s+1}}\,, \tag{4.2}$$

where V is the total variation of $f^{(s)}$. With this stronger assumption $f \in X^s$ half an order is gained by polynomial approximation in contrast to the H^s bases. This is expected as we require half an order of smoothness in L_2 more, cf. Remark 4.4. (In the later numerical experiments, we observe the gain of half an order for H^1 and H^2 as well.)

A drawback comes with the Christoffel function $N(V_m, \cdot)$. Since $N(V_m, 0) = m^2$, cf. (1.1), this spoils the choice of m to quadratic oversampling:

$$10m^2(\log(m) + t) \leq n\,,$$

which is usual for polynomial approximation in uniform points, cf. [31].

- When using the Chebyshev measure $\mathrm{d}\varrho_T(x) = (1-(2x-1)^2)^{-1/2}\,\mathrm{d}x$ we have to compensate with $\beta(x) = (1-(2x-1)^2)^{-1/2}$ to obtain uniform random samples as well. Orthogonalizing the first $m-1$ monomials with respect to the Chebyshev measure, we obtain the Chebyshev polynomials T_k, which are a BOS.

  As for the error, assuming $f \in X^s$, we use [48, Theorem 7.1] or [38, Theorem 6.16] to obtain the same bound as for Legendre polynomials (4.2) but with respect to the $L_2((0,1), (1-(2x-1)^2)^{-1/2}\mathrm{d}x)$ norm.

  As $\|\beta\|_\infty$ diverges at the border, this spoils the choice of the polynomial degree m and our bound. Note, when we exclude some area around the border for sampling, it does not diverge and the resulting error can be controlled. This is called padding and was done in [39, Section 4.1.2]

Thus, with polynomial approximation we assume half an order of smoothness more, cf. Remark 4.4, which we also see in the approximation rate $\mathcal{O}(m^{s+1/2})$.

Remark 4.5. Note, when using the Chebyshev polynomials and samples with respect to the Chebyshev measure, we have $\beta \equiv 1$. Since the Chebyshev polynomials are a BOS, this does not spoil our bounds.

Furthermore, using the Legendre polynomials ($\mathrm{d}\varrho_T = \mathrm{d}x$) and samples with respect to the Chebyshev measure ($\beta(x) = \pi(1-(2x-1)^2)^{1/2}$) works as well. To see this we use [42, Lemma 5.1]:

$$\sqrt{1-(2x-1)^2}|P_k(x)|^2 \leq \frac{2}{\pi}\Big(2 + \frac{1}{k}\Big)$$

for $k \geq 1$. Thus, $\|\beta(\cdot)N(V_m, \cdot)\|_\infty$ and $\|\beta(\cdot)\|_\infty$ are bounded and do not spoil the choice of polynomial degree m nor the error bound.

4.3. **Numerics on the unit interval**

To support our findings, we give a numerical example. As a test function we use

$$f(x) = B_2^{\mathrm{cut}}(x) \quad \text{with} \quad B_2^{\mathrm{cut}}(x) = \begin{cases} -x^2 + 3/4 & \text{for } x \in [0, 1/2] \\ x^2/2 - 3/2x + 9/8 & \text{for } x \in [1/2, 1] \end{cases} \tag{4.3}$$

which was already considered in [40, 35]. The function B_2^{cut} is shown in Figure 1.1 and is a cutout of the B-spline of order two. It and its first derivative are absolute continuous and the second derivative is of bounded variation. Therefore $f \in X^3$ and the polynomial approximation bounds from above are

applicable. According to Remark 4.4 we further have $f \in H^{5/2-\varepsilon}$ for any $\varepsilon > 0$, i.e., there exists $C \geq 0$ such that for $k \geq 0$ it holds $\langle f, \eta_k \rangle_{L_2} \leq Ck^{-5/2+\varepsilon}$. In particular, $f \in H^2$ and (4.1) is applicable for approximating with the H^1 and H^2 bases.

We sample f in 10 000 uniformly random points and add $0.1\%M$ Gaussian noise to obtain $\boldsymbol{y} = \boldsymbol{f} + \boldsymbol{\varepsilon}$, where $M = \max_{x\in[0,1]} f(x) - \min_{x\in[0,1]} f(x) = 5/8$. For V_m we consider the four choices from above: The Chebyshev polynomials $V_m = \operatorname{span}\{T_k\}_{k=0}^{m-1}$ ($\mathrm{d}\varrho_T(x) = (1-(2x-1)^2)^{-1/2}\,\mathrm{d}x$ and $\beta = \pi/2$); the Legendre polynomials $V_m = \operatorname{span}\{P_k\}_{k=0}^{m-1}$ ($\mathrm{d}\varrho_T(x) = \mathrm{d}x$ and $\beta \equiv 1$); the first m basis functions of H^1 from Theorem 4.1, and the the first m basis functions of H^2 from Theorems 4.2 and 4.3 ($\mathrm{d}\varrho_T(x) = \mathrm{d}x$ and $\beta \equiv 1$ as well). For $m = \dim(V_m)$ up to 1 000 we do the following:

(i) Compute the minimal and maximal singular values of $1/\sqrt{n}\boldsymbol{W}^{1/2}\boldsymbol{L}$, with $\boldsymbol{W}$ and $\boldsymbol{L}$ given in (3.1).

(ii) We use least squares with 20 iterations to obtain the approximation $S_m\boldsymbol{y} = \sum_{k=0}^{m-1} \hat{g}_k\eta_k$, defined in (3.1).

(iii) We compute the L_2-error by using Parseval's equality:

$$\|f - S_m\boldsymbol{y}\|_{L_2}^2 = \|f\|_{L_2}^2 - \sum_{k=0}^{m-1} |\hat{f}_k|^2 + \sum_{k=0}^{m-1} |\hat{f}_k - \hat{g}_k|^2\,,$$

where the coefficients $\hat{f}_k = \langle f, \eta_k\rangle_{L_2}$ are computed analyticaly.

(iv) We compute the split approximation error:

$$\|f - S_m\boldsymbol{y}\|_{L_2}^2 \leq 2\|f - S_m\boldsymbol{f}\|_{L_2}^2 + 2\|S_m\boldsymbol{\varepsilon}\|_{L_2}^2\,,$$

where we compute both quantities separately, again, using Parseval's equality.

The results are depicted in Figure 4.1.

- The smallest singular values for the Chebyshev polynomials and the Legendre polynomials decay rapidly for bigger m. This coincides with the violation of the assumtion in Lemma 3.1 for small m:

$$10\|\beta(\cdot)N(V_m,\cdot)\|_\infty(\log(m)+t) \leq n,$$

where $\|\beta(\cdot)N(V_m,\cdot)\|_\infty$ is unbounded in the Chebyshev case and quadratic in the Legendre case, cf. (1.1). In this experiment, for $m = 1\,000$ the condition number $\sigma_{\max}(\boldsymbol{W}^{1/2}\boldsymbol{L})/\sigma_{\min}(\boldsymbol{W}^{1/2}\boldsymbol{L})$ exceeded 10^{29} for the algebraic polynomials and was below 14 for the H^s basis.

- The error for exact function values $\|f - S_m\boldsymbol{f}\|_{L_2}^2$ has decay 3/2 for H^1 and 5/2 for the other bases. This conforms with the theory for the polynomial bases. For the H^1 and H^2 bases the theory predicted only decay rate 1 and 2, cf. Theorems 4.1, 4.2, and (4.1).

- For the noise error $\|S_m\boldsymbol{\varepsilon}\|_{L_2}^2$ we observe linear growth in $m = \dim(V_m)$ as predicted in Theorem 1.1. Furthermore, this error is bigger by a factor of around 40 in the Chebyshev case compared to the others. The maximal weight $\|\boldsymbol{W}\|_\infty$ in this case is around 40 as well. The error due to noise in our bound has the factor $\|\beta\|_\infty$ which can be replaced by $\|\boldsymbol{W}\|_\infty$ to sharpen the bound and explain this effect.

This numerical experiment and the earlier theoretical discussion shows, that the H^1 and the H^2 bases are suitable for approximating functions on the unit interval given in uniform random samples. They are numerically stable in contrast to polynomial approximation with Chebyshev or Legendre. In particular, the least squares matrix is well-conditioned and we can limit the iterations when using an iterative solver, cf. [19, Theorem 3.1.1].

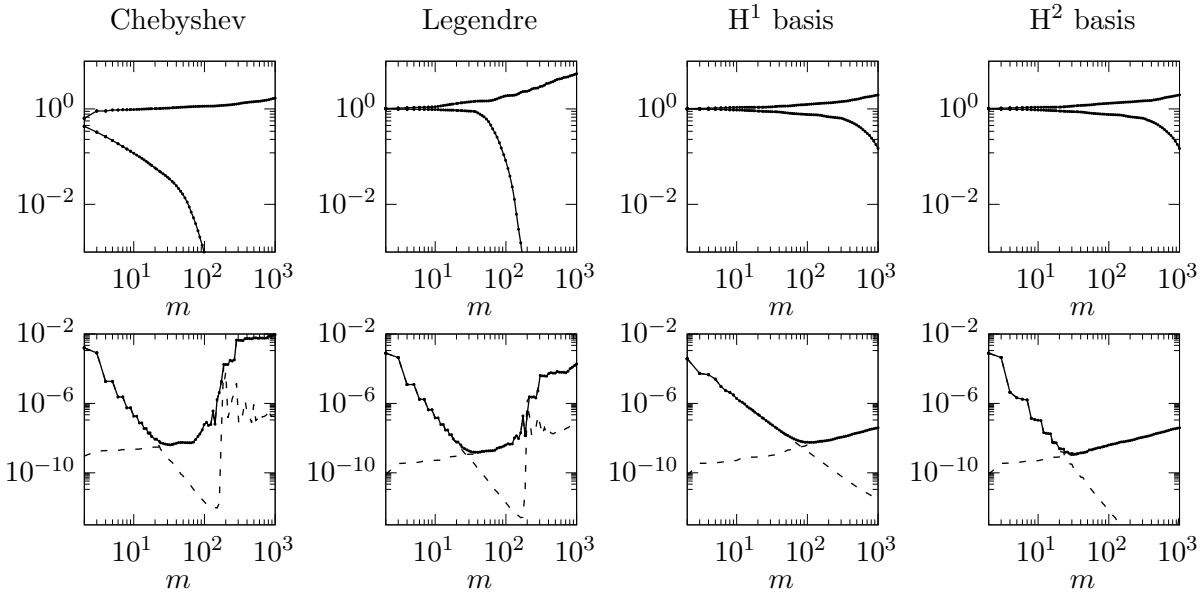

FIGURE 4.1. One-dimensional experiment for different choices of V_m. Top row: minimal and maximal singular value of $1/\sqrt{n}\boldsymbol{W}^{1/2}\boldsymbol{L}$. Bottom row: the L_2-approximation error $\|f - S_m\boldsymbol{y}\|_{L_2}^2$ (solid line) split into the error for exact function values $\|f - S_m\boldsymbol{f}\|_{L_2}^2$ and the noise error $\|S_m\boldsymbol{\varepsilon}\|_{L_2}^2$ (dashed lines) with respect to m.

4.4. **Sobolev spaces with dominating mixed smoothness on the unit cube**

The ideas from Subsection 4.1 can be extended to higher dimensions using the concept of dominating mixed smoothness. We focus on the case of H^1 and H^2, but the same can be done for polynomials as well, cf. [44, Section 8.5.1].

Let $D = [0,1]^d$ be the d-dimensional unit cube equipped with the Lebesgue measure $\mathrm{d}\boldsymbol{x}$. The Sobolev space with dominating mixed smoothness of integer degree $s \geq 0$ is given by $H^s_{\mathrm{mix}} = H^s_{\mathrm{mix}}(0,1)^d = H^s(0,1) \otimes \cdots \otimes H^s(0,1)$. The inner product of these Hilbert spaces is given by

$$\langle f, g\rangle_{H^s_{\mathrm{mix}}} = \sum_{\boldsymbol{j}\in\{0,s\}^d} \langle D^{(\boldsymbol{j})} f, D^{(\boldsymbol{j})} g\rangle_{L_2}\,.$$

With σ_k and η_k the singular values and eigenfunctions of H^s, the singular values and eigenfunctions of $W = \mathrm{Id}^* \circ \mathrm{Id}\colon H^s_{\mathrm{mix}} \to H^s_{\mathrm{mix}}$ extend as follows:

$$\sigma^2_{\boldsymbol{k}} = \prod_{j=1}^d \sigma^2_{k_j} \quad \text{and} \quad \eta_{\boldsymbol{k}}(\boldsymbol{x}) = \prod_{j=1}^d \eta_{k_j}(x_j)\,.$$

To obtain the eigenfunctions corresponding to the smallest singular values, we now work with multi-indices $\boldsymbol{k}$. The indices corresponding to the largest singular values lie on a, so called, hyperbolic cross

$$I_R(H^s_{\mathrm{mix}}) \coloneqq \Big\{\boldsymbol{k} \in \mathbb{N}^d : \prod_{j=1}^d \sigma^2_{k_j} \geq R\Big\}\,.$$

For $V_m = \operatorname{span}\{\eta_{\boldsymbol{k}} : \boldsymbol{k} \in I_R(H^s_{\mathrm{mix}})\}$ and $f \in H^s_{\mathrm{mix}}$, we obtain by (4.1)

$$e(f, V_m, L_2)_{L_2} \leq R\|f\|^2_{H^s_{\mathrm{mix}}}\,.$$

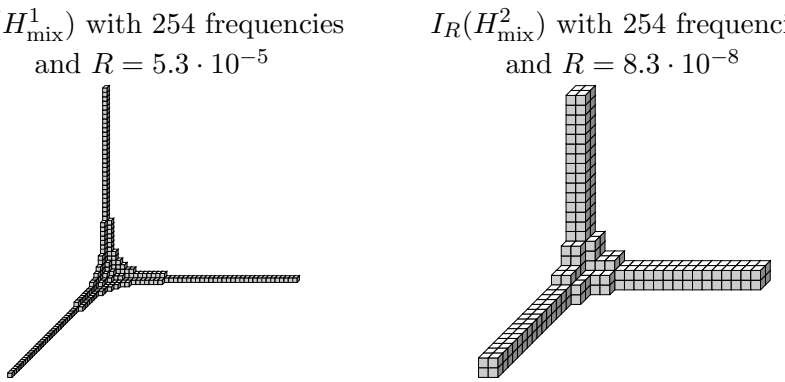

FIGURE 4.2. Hyperbolic cross in three dimensions.

In Figure 4.2 we have equally sized index sets for H^1_{mix} and H^2_{mix}. Note, that R is smaller for H^2_{mix} as its singular values decay faster, cf. Theorems 4.1 and 4.2.

4.5. **Numerics on the unit cube**

For a numerical experiment we do the same as in the one-dimensional case but only consider the H^2_{mix} case. For our test function we tensorize the B-Spline cutout

$$f(\boldsymbol{x}) = \prod_{j=1}^{d} B_2^{\text{cut}}(x_j)$$

where B_2^{cut} was defined in (4.3). We increase the dimension to $d = 5$ and the number of samples to 1 000 000 and use Gaussian noise with variance $\sigma^2 \in \{0.00, 0.01M, 0.03M\}$ where $M = \max_{\boldsymbol{x}\in[0,1]^d} f(\boldsymbol{x}) - \min_{\boldsymbol{x}\in[0,1]^d} f(\boldsymbol{x}) = 5/8$.

Let $V_m = \operatorname{span}\{\eta_{\boldsymbol{k}} : \boldsymbol{k} \in I_R(H^2_{\text{mix}})\}$ of size m with $\eta_{\boldsymbol{k}}$ the tensorized H^2_{mix} basis, cf. Theorems 4.2 and 4.3. Since the H^2_{mix} basis is a BOS, we obtain

$$\frac{N(V_m)}{m} \leq 6\,.$$

With $t = 6$, we satisfy the assumptions of Theorem 1.1 for $m \leq 12\,250$ and obtain a probability exceeding 0.99 for the error bound in Theorem 1.1. For $m = \dim(V_m)$ up to 10 000 we do the following:

(i) We use plain least squares with 20 iterations to obtain the approximation $S_m\boldsymbol{y} = \sum_{k=0}^{m-1} \hat{g}_k \eta_k$, defined in (3.1).

(ii) We compute the L_2-error by using Parseval's equality analog to the one-dimensional case.

(iii) We compute our bound: Applying Theorem 1.1 using (3.5), $t = 6$, and $n = 1\,000\,000$, we obtain

$$\begin{aligned}\|f - S_m\boldsymbol{y}\|_{L_2}^2 \leq 14\Big(1 + \sqrt{\frac{6N(V_m)}{n}}\Big)^2 e(f, V_m, L_2)_{L_2}^2 \\ + \frac{m}{n}\Big(138B\sqrt{\sigma^2} + 4\sigma^2\Big) + 0.0031B^2\end{aligned}$$

with probability exceeding 0.99 and all the remaining quantities known in our experiment.

The results are depicted in Figure 4.3.

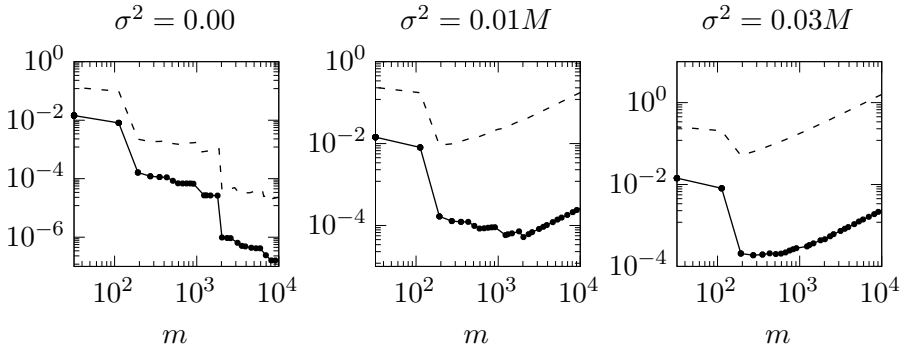

FIGURE 4.3. Five-dimensional experiment for H^2_{mix}. The solid lines represent the L_2-error $\|f - S_m \boldsymbol{y}\|^2_{L_2}$ and the dashed lines the bound from Theorems 3.2 and 1.1.

- The bounds capture the error behaviour well. But it seems that there is room for improvement in the constants, especially in the experiments with noise. Here, improving constants in the Hanson-Wright inequality in Theorem 2.3 could be a starting point.

- Furthermor, this experiment shows, that the H^2 basis is easily suitable for high-dimensional approximation as well.

Acknowledgement

I would like to thank Tino Ullrich for many helpful discussions and his expertise with function spaces. Further, I would like to thank Michael Schmischke for several insights on approximation on the unit cube, Sergei Pereverzyev for his knowledge on regularization theory as well as Ralf Hielscher and Daniel Potts for further discussions.

Appendix A. Calculations for the $H^2(0,1)$ basis

Proof. [Proof of the first part of Theorem 4.2] Analogously to Theorem 4.1, for σ a singular value of W with corresponding eigenfunction $\eta \in H^2_{\text{mix}}$, we obtain the following differential equation

$$\frac{1-\sigma^2}{\sigma^2}\eta = \eta^{(4)} \quad \text{with} \quad \eta^{(2)}(0) = \eta^{(2)}(1) = \eta^{(3)}(0) = \eta^{(3)}(1) = 0.$$

Now we distinguish three cases for the value of σ^2:
First case. Let us assume $\sigma^2 = 1$. The ansatz function becomes

$$\eta(x) = A + Bx + Cx^2 + Dx^3.$$

From the conditions $\eta^{(2)}(0) = \eta^{(3)}(0) = 0$ we obtain $D = C = 0$. The two remaining degrees of freedom are restricted by demanding $L_2(0,1)$-orthonormality. By simple calculus we obtain the proposed eigenfunctions η_0 and η_1.
Second case. Lets assume $(1-\sigma^2)/\sigma^2 > 0 \Leftrightarrow \sigma^2 < 1$. Introducing $t := \sqrt[4]{(1-\sigma^2)/\sigma^2}$, we use the ansatz

$$\eta(x) = A\cos(tx) + B\sin(tx) + C\cosh(tx) + D\sinh(tx).$$

The conditions $\eta^{(2)}(0) = \eta^{(3)}(0) = 0$ transform to $A = C$ and $B = D$, respectively. The conditions $\eta^{(2)}(1) = \eta^{(3)}(1) = 0$ can be put into a system of equations:

$$\begin{bmatrix} \cosh(t) - \cos(t) & \sinh(t) - \sin(t) \\ \sinh(t) + \sin(t) & \cosh(t) - \cos(t) \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} = \mathbf{0}$$

or, by using $\cosh^2(t) - \sinh^2(t) = \cos^2(t) + \sin^2(t) = 1$, equivalently

$$\begin{bmatrix} \cosh(t) - \cos(t) & \sinh(t) - \sin(t) \\ 0 & 1 - \cosh(t)\cos(t) \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} = \mathbf{0}.$$

For non-trivial solutions we need non-regularity of that matrix which transforms to the condition $\cosh(t)\cos(t) = 1$. With the leftover degree of freedom we choose

$$A = C = 1 \quad \text{and} \quad B = D = -\frac{\cosh(t) - \cos(t)}{\sinh(t) - \sin(t)}$$

and obtain η_k for $k \geq 2$ as proposed in the theorem.
For the L_2-norm we obtain

$$\begin{aligned}
\int_0^1 |\eta_n|^2 \, dx \quad &= \int_0^1 (\cosh(tx) + \cos(tx))^2 \, dx + B^2 \int_0^1 (\sinh(tx) + \sin(tx))^2 \, dx \\
&\quad + 2B \int_0^1 (\cosh(tx) + \cos(tx))(\sinh(tx) + \sin(tx)) \, dx \\
&= 1 + \frac{\sin(2t) + \sinh(2t) + 4\cos(t)\sinh(t) + 4\sin(t)\cosh(t)}{4t} \\
&\quad + B^2 \frac{-\sin(2t) + \sinh(2t) - 4\cos(t)\sinh(t) + 4\sin(t)\cosh(t)}{4t} + 2B\frac{(\sin(t) + \sinh(t))^2}{2t} \\
&= 1 + \frac{1 + B^2}{4t}(\sinh(2t) + 4\sin(t)\cosh(t)) + \frac{1 - B^2}{4t}(\sin(2t) + 4\cos(t)\sinh(t)) \\
&\quad + B\frac{(\sin(t) + \sinh(t))^2}{t} \\
&= 1 + \frac{1 + B^2}{2t}\cosh(t)(\sinh(t) + 2\sin(t)) + \frac{1 - B^2}{2t}\cos(t)(\sin(t) + 2\sinh(t)) \\
&\quad + B\frac{(\sin(t) + \sinh(t))^2}{t}.
\end{aligned}$$

Using $\cos(t)\cosh(t) = 1$, we obtain

$$1 + B^2 = 2\frac{\sinh(t)}{\sinh(t) - \sin(t)} \quad \text{and} \quad 1 - B^2 = -2\frac{\sin(t)}{\sinh(t) - \sin(t)}. \tag{A.1}$$

Thus,

$$\begin{aligned}
\int_0^1 |\eta_n|^2 \, dx &= 1 + \frac{1}{t(\sinh(t) - \sin(t))} \Big[\sinh(t)\cosh(t)(\sinh(t) + 2\sin(t)) \\
&\qquad - \sin(t)\cos(t)(\sin(t) + 2\sinh(t)) - (\cosh(t) - \cos(t))(\sin(t) + \sinh(t))^2 \Big] \\
&= 1 + \frac{\cos(t)\sinh^2(t) - \cosh(t)\sin^2(t)}{t(\sinh(t) - \sin(t))} \\
&= 1 + \frac{\cos(t)\cosh^2(t) - \cos(t) - \cosh(t) + \cosh(t)\cos^2(t)}{t(\sinh(t) - \sin(t))}
\end{aligned}$$

where $\cos^2(t) + \sin^2(t) = \cosh^2(t) - \sinh^2(t) = 1$ was used in the last equality. Using $\cosh(t)\cos(t) = 1$, the latter summand evaluates to zero and we have proven the L_2-normality.

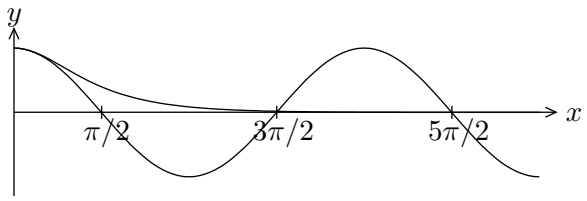

FIGURE A.1. $\cos(t)$ and $1/\cosh(t)$

Third case. Assume $\sigma^2 > 1$. Set $t := \sqrt[4]{(\sigma^2-1)/(\sigma^2)}$. The ansatz becomes

$$\begin{aligned}\eta(x) = A\cosh(tx)\cos(tx) + B\cosh(tx)\sin(tx)\\ + C\sinh(tx)\cos(tx) + D\sinh(tx)\sin(tx).\end{aligned}$$

The conditions $\eta^{(2)}(0) = \eta^{(3)}(0) = 0$ transform to $D = 0$ and $B = C$. The two remaining degrees of freedom are fixed by the conditions $\eta^{(2)}(1) = \eta^{(3)}(1) = 0$ which, in matrix form, look as follows

$$\begin{bmatrix} -\sinh(t)\sin(t) & \sinh(t)\cos(t) - \cosh(t)\sin(t) \\ -\sinh(t)\cos(t) - \cosh(t)\sin(t) & -2\sinh(t)\sin(t) \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

For a non-trivial solution we need that matrix to be non-regular. To achieve that we have a look at the roots of its determinant:

$$2\sinh^2(t)\sin^2(t) + \sinh^2(t)\cos^2(t) - \cosh^2(t)\sin^2(t) \stackrel{!}{=} 0.$$

Using $\sin^2(t) + \cos^2(t) = \cosh^2(t) - \sinh^2(t) = 1$ we have

$$\sinh^2(t) - \sin^2(t) = \frac{1}{2}\cosh(2t) + \frac{1}{2}\cos(2t) - 1 \stackrel{!}{=} 0$$

which is only fulfilled for $t = 0$, or equivalently, $\sigma^2 = 1$. Hence, there are no eigenvalues bigger than 1. ∎

Lemma A.1. *For $0 < t_2 < t_3 < \dots$ fulfilling $\cosh(t_k)\cos(t_k) = 1$ and $\tilde{t}_k = \frac{2k-1}{2}\pi$, we have*

$$\frac{3}{2}\pi < t_2 \quad \textit{and} \quad \left|\tilde{t}_k - t_k\right| \le \varepsilon$$

for $k \ge \frac{1}{\pi}\log(\pi/\varepsilon)$. In particular $|\tilde{t}_k - t_k| \le \pi\exp(-2\pi)$ for all $k \ge 2$.

Proof.
Since $0 < 1/\cosh(t) < 1$ for $t > 0$ and the oscillating behaviour of $\cos(t)$, as depicted in Figure A.1, we obtain

$$t_k \in \begin{cases} \left(\frac{2k-1}{2}\pi, \frac{2k}{2}\pi\right) & \text{for } k \text{ even} \\ \left(\frac{2k-2}{2}\pi, \frac{2k-1}{2}\pi\right) & \text{for } k \text{ odd}. \end{cases}$$

In particular, $\frac{3}{2}\pi < t_2$. Furthermore, for even k and $t \in \left(\frac{2k-1}{2}\pi, \frac{2k}{2}\pi\right)$ we have

$$\frac{1}{\cosh(t)} \le 2\exp(-t) \le 2\exp\Big(-\frac{2k-1}{2}\pi\Big) \quad \text{and} \quad \cos(t) \ge \frac{t - \frac{2k-1}{2}\pi}{\pi/2}.$$

The function bounds intersect for a value larger than t_k, which we use to refine the interval:

$$t_k \in \Big(\tilde{t}_k, \tilde{t}_k + \pi\exp\Big(-\frac{2k-1}{2}\pi\Big)\Big).$$

Similarly, for odd k and $t \in \left(\frac{2k-2}{2}\pi, \frac{2k-1}{2}\pi\right)$ we obtain

$$t_k \in \left(\tilde{t}_k - \pi \exp\Big(-\frac{2k-2}{2}\pi\Big), \tilde{t}_k\right).$$

Thus, for $k \geq 2$ we have $\left|\tilde{t}_k - t_k\right| \leq \pi \exp(-(k-1)\pi)$, which is smaller than ε for

$$k \geq \frac{\log(\pi/\varepsilon)}{\pi} + 1 \geq \frac{\log(\pi/\varepsilon)}{\pi}.$$

■

Lemma A.2. *For $0 < t_2 < t_3 < \dots$ fulfilling $\cosh(t_k)\cos(t_k) = 1$, we have that η_{t_k} defined by*

$$\eta^I_{t_k}(x) = \cosh(t_k x) - \frac{\cosh(t_k) - \cos(t_k)}{\sinh(t_k) - \sin(t_k)} \sinh(t_k x) \tag{A.2}$$

is convex and non-negative for all even k and monotone for all odd k.

Proof. Step 1. We distinguish for different values of $B = B(t) \coloneqq (\cosh(t) - \cos(t))/(\sinh(t) - \sin(t))$. For $B < 1$ we have

$$\eta^I_t(x) = \cosh(tx) - B\sinh(tx) \geq \cosh(tx) - \sinh(tx) \geq 0$$

and by the same argument $(\eta^I_t(x))^{(2)} = t^2 \eta^I_t(x) \geq 0$ for all $x \geq 0$. Thus, $\eta^I_t(x)$ is convex and non-negative.
For $B > 1$ we obtain

$$(\eta^I_t)' = t(\sinh(tx) - B\cosh(tx)) \leq t(\sinh(tx) - \cosh(tx)) \leq 0$$

for all $x \geq 0$. Thus, η^I_t is monotone.
Step 2. It is left to show for which k's $B(t_k)$ attains a value smaller or bigger than one:

$$\begin{aligned} B(t_k) \lesseqgtr 1 \quad &\Leftrightarrow \quad \cosh(t_k) - \cos(t_k) \lesseqgtr \sinh(t_k) - \sin(t_k) \\ &\Leftrightarrow \quad \exp(-t_k) - \sqrt{2}\cos(t_k + \pi/4) \lesseqgtr 0. \end{aligned}$$

We will show that $\exp(-t_k) - \sqrt{2}\cos(t_k + \pi/4)$ has the same sign as $(-1)^{k+1}$ and, thus, are finished. We do this by estimating their difference by a quantity smaller than one. With $\tilde{t}_k = \frac{2k-1}{2}\pi$ we obtain

$$\begin{aligned} &|\exp(-t_k) - \sqrt{2}\cos(t_k + \pi/4) - (-1)^{k+1}| \\ &\quad = |\exp(-t_k) - \sqrt{2}\cos(t_k + \pi/4) + \sqrt{2}\cos(\tilde{t}_k + \pi/4)|. \end{aligned}$$

Using that cos is Lipschitz-continuous with constant 1 and Lemma A.1 we estimate the above by

$$\begin{aligned} |\exp(-t_k) - \sqrt{2}\cos(t_k + \pi/4) - (-1)^{k+1}| &\leq |\exp(-t_k)| + \sqrt{2}|t_k - \tilde{t}_k| \\ &\leq \exp(-3/2\pi) + \sqrt{2}\pi \exp\Big(-2\pi\Big), \end{aligned}$$

which is certainly smaller than one. ■

Lemma A.3. *For $0 < t_2 < t_3 < \dots$ fulfilling $\cosh(t_k)\cos(t_k) = 1$, we have that $\eta^I_{t_k}$ defined in* (A.2) *is even with respect to the axis $x = 1/2$ for all even k and vice versa.*

Proof. Step 1. We will show that $\eta^I_{t_k}$ has any symmetry around $x = 1/2$. We shift the function and split it into an odd and an even part. For $B = B(t) = (\cosh(t) - \cos(t))/(\sinh(t) - \sin(h))$, we obtain

$$\begin{aligned} \eta^I_t(x + 1/2) &= \cosh(tx + t/2) - B\sinh(tx + t/2) \\ &= \underbrace{(\cosh(t/2) - B\sinh(t/2))}_{=:\alpha}\cosh(tx) + \underbrace{(\sinh(t/2) - B\cosh(t/2))}_{=:\beta}\sinh(tx). \end{aligned}$$

Multiplying the two factors α and β in front of $\cosh(tx)$ and $\sinh(tx)$, we obtain

$$\begin{aligned}\alpha \cdot \beta &= -B\cosh^2(t/2) - B\sinh^2(t/2) + (1+B^2)\cosh(t/2)\sinh(t/2)\\ &= -B\frac{\cosh(t)-1}{2} - B\frac{\cosh(t)+1}{2} + (1+B^2)\frac{\sinh(t)}{2}\\ &= -B\cosh(t) + (1+B^2)\frac{\sinh(t)}{2}\,.\end{aligned}$$

Using (A.1), $\cosh(t)\cos(t) = 1$, and $1 = \cosh^2(t) - \sinh^2(t)$ this evaluates to

$$\alpha \cdot \beta = -\frac{\cosh^2(t)-1}{\sinh(t)-\sin(t)} + \frac{\sinh^2(t)}{\sinh(t)-\sin(t)} = 0\,.$$

And since we are not dealing with the zero function either α or β is zero. Thus, $x \mapsto \eta_t^I(x+1/2)$ obeys a symmetry.

Step 2. It remains to specify the kind of symmetry. By Lemma A.2 we have that η_t^I is convex for even k. Since a convex non-constant function cannot be odd it has to be even. Also by Lemma A.2 we have that η_t^I is monotone for odd k. Since a monotone non-zero function cannot be even it has to be odd. ∎

Proof. [Proof of the second part of Theorem 4.2] The cases $k \in \{0,1\}$ are clear. For $k \geq 2$ we split the function into η_t^I defined in (A.2) and

$$\eta_t^{II}(x) = \cos(tx) - \frac{\cosh(t)-\cos(t)}{\sinh(t)-\sin(t)}\sin(tx).$$

We will show that each of these is bounded by $1.01\sqrt{2}$ and, thus, obtain the assertion.

Step 1. In order to bound $\eta_{t_k}^I$ we firstly have a look at the boundary points $x \in \{0,1\}$. With Lemma A.3 we obtain

$$\eta_{t_k}^I(0) = \left|\eta_{t_k}^I(1)\right| = 1. \tag{A.3}$$

By Lemma A.2 $\eta_{t_k}^I$ is either non-negative and convex or monotone and, thus, cannot exceed its values on the boundary.

Step 2. In order to bound η_t^{II} we define

$$B \coloneqq \frac{\cosh(t)-\cos(t)}{\sinh(t)-\sin(t)} \quad \text{and} \quad \vartheta = \arg(1+Bi).$$

Next, we use the exponential definition of sine and cosine and the polar representation of complex numbers to obtain

$$\begin{aligned}\eta_t^{II}(x) &= \cos(tx) - B\sin(tx)\\ &= \frac{\exp(itx)+\exp(-itx)}{2} + Bi\frac{\exp(itx)-\exp(-itx)}{2}\\ &= \frac{(1+Bi)\exp(itx)+(1-Bi)\exp(-itx)}{2}\\ &= \sqrt{1+B^2}\frac{\exp(i(tx+\vartheta))+\exp(-i(tx+\vartheta))}{2}\\ &= \sqrt{1+B^2}\cos(tx+\vartheta)\end{aligned}$$

Thus, by (A.1)

$$|\eta_t^{II}(x)| \leq \sqrt{1+B^2} = \sqrt{\frac{2}{1-\sin(t)/\sinh(t)}} \leq \sqrt{\frac{2}{1-1/\sinh(t)}} \tag{A.4}$$

From Lemma A.1 we use $t \geq 3/2\pi$ in combination with the monotonicity in (A.4) we have $|\eta_t^{II}(x)| \leq 1.01\sqrt{2}$.

■

Lemma A.4. *For $t \geq \max\{2\log(4/\varepsilon), 3/2\pi\}$ we have for $x \in [0, 1/2]$*

$$\Big|\Big(1 - \frac{\cosh(t) - \cos(t)}{\sinh(t) - \sin(t)}\Big)\sinh(tx)\Big| \leq \varepsilon\,.$$

Proof. We use $\cosh(t) - \sinh(t) = \exp(-t)$ and $\cos(t) - \sin(t) = \sqrt{2}\cos(t + \pi/4)$ to estimate

$$\Big|\Big(1 - \frac{\cosh(t) - \cos(t)}{\sinh(t) - \sin(t)}\Big)\sinh(tx)\Big| = |\sqrt{2}\cos(t + \pi/4) - \exp(-t)|\Big|\frac{\sinh(tx)}{\sinh(t) - \sin(t)}\Big|.$$

Since $x \leq 1/2$, sinh strictly monotone growing, and $t \geq 3/2\pi$ by Lemma A.1, we further estimate

$$\begin{aligned}\Big|\Big(1 - \frac{\cosh(t) - \cos(t)}{\sinh(t) - \sin(t)}\Big)\sinh(tx)\Big| &\leq 2\Big|\frac{\sinh(t/2)}{\sinh(t) - \sin(t)}\Big| \\ &= 2\Big|\frac{1}{2\cosh(t/2)}\frac{1}{1 - \sin(t)/\sinh(t)}\Big|\end{aligned}$$

Using $1 - \sin(t)/\sinh(t) > 1/2$ for $t > 3/2\pi$, we obtain

$$\Big|\Big(1 - \frac{\cosh(t) - \cos(t)}{\sinh(t) - \sin(t)}\Big)\sinh(tx)\Big| \leq \frac{2}{\cosh(t/2)} \leq \frac{4}{\exp(t/2)}\,,$$

which is smaller than ε for $t \geq 2\log(4/\varepsilon)$. ■

Lemma A.5. *For $0 < t_2 < t_3 < \ldots$ fulfilling $\cosh(t_k)\cos(t_k) = 1$, we have*

$$\Big|\eta_{t_k}^{II}(x) - \sqrt{2}\cos(t_k x + \pi/4)\Big| \leq \varepsilon \quad \textit{for} \quad x \in [0, 1]$$

and

$$\Big|\eta_{t_k}^{I}(x) - \exp(-tx)\Big| \leq \varepsilon \quad \textit{for} \quad x \in [0, 1/2]$$

for $k \geq \frac{2}{\pi}\log(4/\varepsilon) + 1$.

Proof. Step 1. For the first inequality we use

$$\sqrt{2}\cos(tx + \pi/4) = \cos(tx) - \sin(tx)$$

to obtain

$$\Big|\eta_t^{II} - \sqrt{2}\cos(tx + \pi/4)\Big| = \Big|\Big(1 - \frac{\cosh(t) - \cos(t)}{\sinh(t) - \sin(t)}\Big)\sin(tx)\Big|$$

which is smaller than ε for $t > \max\{2\log(4/\varepsilon), 3/2\pi\}$ by Lemma A.4.
The second inequality follows analogously from $\exp(-tx) = \cosh(tx) - \sinh(tx)$ and Lemma A.4.
Step 2. It is left to show the condition $t \geq \max\{2\log(4/\varepsilon), 3/2\pi\}$ from Step 1. By Lemma A.1 we have $t_k \geq 3/2\pi$. Further, by assumption, we have

$$k \geq \frac{2}{\pi}\log\Big(\frac{4}{\varepsilon}\Big) + 1 \geq \frac{2}{\pi}\log\Big(\frac{4}{\varepsilon}\Big) + \exp(-2\pi) + \frac{1}{2}$$

Thus,

$$2\log\Big(\frac{4}{\varepsilon}\Big) \leq \frac{2k - 1}{2}\pi - \pi\exp(-2\pi) \leq t_k$$

where the last inequality follows from Lemma A.1. ■

Proof. [Proof of Theorem 4.3] Because of the symmetry shown in Lemma A.3 we assume without loss of generality $x \in [0, 1/2]$. Then

$$|\eta_n(x) - \tilde{\eta}_n(x)| \leq \left|\eta_n^I(x) - \exp(-t_k x)\right| + \left|\eta_n^{II}(x) - \sqrt{2}\cos(t_k x + \pi/4)\right| + \left|\exp(-t_k x) - \exp(-\tilde{t}_k x)\right| + \left|\sqrt{2}\cos(t_k x + \pi/4) - \sqrt{2}\cos(\tilde{t}_k x + \pi/4)\right|.$$

By Lemma A.5, the first two summands are each smaller than $\varepsilon/4$ each for $n > \frac{2}{\pi}\log(16/\varepsilon) + 1$. We estimate the two latter summands as follows.
Since cos is Lipschitz continuous with constant one we have

$$\left|\sqrt{2}\cos(t_k x + \pi/4) - \sqrt{2}\cos\left(\tilde{t}_k x + \frac{\pi}{4}\right)\right| \leq \left|\sqrt{2}\left(t_k - \tilde{t}_k\right)\right|$$

which, by Lemma A.1 is smaller than $\varepsilon/4$ for $k > \frac{1}{\pi}\log(4\pi/(\sqrt{2}\varepsilon))$.
Since exp is Lipschitz continuous with constant 1 on $(-\infty, 0)$, we have

$$\left|\exp(-t_k x) - \exp(-\tilde{t}_k x)\right| \leq \left|t_k - \tilde{t}_k\right|$$

which, by Lemma A.1 is smaller than $\varepsilon/4$ for $k > \frac{1}{\pi}\log(16/\varepsilon)$.
Overall, we obtain $|\eta_n(x) - \tilde{\eta}_n(x)| < 4\frac{\varepsilon}{4} = \varepsilon$ for

$$k > \max\left\{\frac{2}{\pi}\log(16/\varepsilon) + 1,\ \frac{1}{\pi}\log(4\pi/(\sqrt{2}\varepsilon)),\ \frac{1}{\pi}\log(16/\varepsilon)\right\} = \frac{2}{\pi}\log(16/\varepsilon) + 1\,.$$

■

Bibliography

[1] B. Adcock. Multivariate modified Fourier series and application to boundary value problems. *Numer. Math.*, 115(4):511–552, 2010.

[2] B. Adcock and D. Huybrechs. Multivariate modified Fourier expansions. In *Spectral and high order methods for partial differential equations*, volume 76 of *Lect. Notes Comput. Sci. Eng.*, pages 85–92. Springer, Heidelberg, 2011.

[3] B. Adcock, A. Iserles, and S. P. Nørsett. From high oscillation to rapid approximation II: expansions in Birkhoff series. *IMA J. Numer. Anal.*, 32(1):105–140, 2012.

[4] Pierre C. B. Concentration of quadratic forms under a Bernstein moment assumption. *arXiv:1901.08736*, 2019.

[5] Y. Baraud. Model selection for regression on a random design. *ESAIM Probab. Stat.*, 6:127–146, 2002.

[6] F. Bartel and R. Hielscher. Concentration inequalities for cross-validation in scattered data approximation. *J. Approx. Theory*, 277:105715, 2022.

[7] F. Bartel, R. Hielscher, and D. Potts. Fast cross-validation in harmonic approximation. *Appl. Comput. Harmon. Anal.*, 49(2):415–437, 2020.

[8] F. Bartel, M. Schäfer, and T. Ullrich. Constructive subsampling of finite frames with applications in optimal function recovery. *Appl. Comput. Harmon. Anal.*, to appear, 2023.

[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 Math. Model. Numer. Anal.*, 49(3):815–837, 2015.

[10] 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.

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

[12] R. Cools, F. Y. Kuo, D. Nuyens, and G. Suryanarayana. Tent-transformed lattice rules for integration and approximation of multivariate non-periodic functions. *J. Complexity*, 36:166–181, 2016.

[13] J. Dick, D. Nuyens, and F. Pillichshammer. Lattice rules for nonperiodic smooth integrands. *Numer. Math.*, 126(2):259–291, 2014.

[14] M. Dolbeault and A. Cohen. Optimal pointwise sampling for L^2 approximation. *Journal of Complexity*, 68:101602, February 2022.

[15] M. Dolbeault and A. Cohen. Optimal sampling and Christoffel functions on general domains. *Constr. Approx.*, 56(1):121–163, 2022.

[16] M. Dolbeault, D. Krieg, and M. Ullrich. A sharp upper bound for sampling numbers in L_2. *Appl. Comput. Harmon. Anal.*, 63:113–134, 2023.

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

[18] E. R. Gizewski, L. Mayer, B. A. Moser, D. H. Nguyen, S. Pereverzyev, S. V. Pereverzyev, N. Shepeleva, and W. Zellinger. On a regularization of unsupervised domain adaptation in RKHS. *Applied and Computational Harmonic Analysis*, 57:201–227, 2022.

[19] A. Greenbaum. *Iterative methods for solving linear systems*, volume 17 of *Frontiers in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.

[20] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002.

[21] C. Haberstich, A. Nouy, and G. Perrin. Boosted optimal weighted least-squares. *Math. Comp.*, January 2022.

[22] J. Hampton and A. Doostan. Coherence motivated sampling and convergence analysis of least squares polynomial Chaos regression. *Comput. Methods Appl. Mech. Engrg.*, 290:73–97, 2015.

[23] A. Iserles and S. P. Nørsett. From high oscillation to rapid approximation. I. Modified Fourier expansions. *IMA J. Numer. Anal.*, 28(4):862–887, 2008.

[24] L. Kämmerer, T. Ullrich, and T. Volkmer. Worst-case recovery guarantees for least squares approximation using random samples. *Constr. Approx.*, 54(2):295–352, 2021.

[25] M. G. Krein. On a special class of differential operators. *Doklady AN USSR*, 2:345–349, 1935.

[26] D. Krieg and M. Ullrich. Function values are enough for L_2-approximation. *Found. Comput. Math.*, 21(4):1141–1151, 2021.

[27] F. Y. Kuo, G. Migliorati, F. Nobile, and D. Nuyens. Function integration, reconstruction and approximation using rank-1 lattices. *Math. Comp.*, 90(330):1861–1897, 2021.

[28] I. Limonova and V. N. Temlyakov. On sampling discretization in L_2. *J. Math. Anal. Appl.*, 515(2):Paper No. 126457, 14, 2022.

[29] L. Lippert, D. Potts, and T. Ullrich. Fast hyperbolic wavelet regression meets ANOVA. *Numer. Math.*, to appear.

[30] S. Lu, P. Mathé, and S. V. Pereverzev. Balancing principle in supervised learning for a general regularization scheme. *Appl. Comput. Harmon. Anal.*, 48(1):123–148, 2020.

[31] G. Migliorati, F. Nobile, E. von Schwerin, and R. Tempone. Analysis of discrete L^2 projection on polynomial spaces with random evaluations. *Found. Comput. Math.*, March 2014.

[32] M. Moeller and T. Ullrich. L_2-norm sampling discretization and recovery of functions from RKHS with finite trace. *Sampl. Theory Signal Process. Data Anal.*, 19(2):13, 2021.

[33] N. Nagel, M. Schäfer, and T. Ullrich. A new upper bound for sampling numbers. *Found. Comp. Math.*, April 2021.

[34] A. Narayan, J. D. Jakeman, and T. Zhou. A Christoffel function weighted least squares algorithm for collocation approximations. *Math. Comp.*, 86(306):1913–1947, 2017.

[35] R. Nasdala and D. Potts. A note on transformed Fourier systems for the approximation of non-periodic signals. In *Monte Carlo and quasi-Monte Carlo methods*, volume 387 of *Springer Proc. Math. Stat.*, pages 253–271. Springer, Cham, [2022] ©2022.

[36] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[37] S. V. Pereverzyev and S. Lu. *Regularization Theory for Ill-Posed Problems. Selected Topics.* DeGruyter, Berlin, Boston, 2013.

[38] G. Plonka, D. Potts, G. Steidl, and M. Tasche. *Numerical Fourier analysis.* Applied and Numerical Harmonic Analysis. Birkhäuser/Springer, Cham, 2018.

[39] D. Potts and M. Schmischke. Learning multivariate functions with low-dimensional structures using polynomial bases. *J. Comput. Appl. Math.*, 403:113821, March 2022.

[40] D. Potts and T. Volkmer. Fast and exact reconstruction of arbitrary multivariate algebraic polynomials in Chebyshev form. In *2015 International Conference on Sampling Theory and Applications (SampTA)*, pages 392–396, 2015.

[41] K. Pozharska and T. Ullrich. A note on sampling recovery of multivariate functions in the uniform norm. *SIAM Journal on Numerical Analysis*, 60(3):1363–1384, 2022.

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

[43] M. Rudelson and R. Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18(none), 2013.

[44] J. Shen, T. Tang, and L. Wang. *Spectral methods*, volume 41 of *Springer Series in Computational Mathematics*. Springer, Heidelberg, 2011. Algorithms, analysis and applications.

[45] I. Steinwart and A. Christmann. *Support vector machines*. Information Science and Statistics. Springer, New York, 2008.

[46] G. Suryanarayana, D. Nuyens, and R. Cools. Reconstruction and collocation of a class of non-periodic functions by sampling along tent-transformed rank-1 lattices. *J. Fourier Anal. Appl.*, 22(1):187–214, 2016.

[47] V. N. Temlyakov. On approximate recovery of functions with bounded mixed derivative. *J. Complexity*, 9(1):41–59, 1993. Festschrift for Joseph F. Traub, Part I.

[48] L. N. Trefethen. *Approximation theory and approximation practice*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2013.

[49] H. Triebel. *Theory of function spaces*. Modern Birkhäuser Classics. Birkhäuser Verlag, Basel, 2010.

[50] Hans Triebel. *Theory of function spaces. II*, volume 84 of *Monographs in Mathematics*. Birkhäuser Verlag, Basel, 1992.

[51] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434, 2012.

[52] H. Wang. New error bounds for Legendre approximations of differentiable functions. *arXiv:2111.03833*, 2021.

[53] A. G. Werschulz and H. Woźniakowski. Tractability of multivariate approximation over a weighted unanchored sobolev space. *Constr. Approx.*, 30(3):395–421, 2009.