Advances in Computational Mathematics (2024) 50:53
https://doi.org/10.1007/s10444-024-10147-2

Variable transformations in combination with wavelets and ANOVA for high-dimensional approximation

Daniel Potts[1] · Laura Weidensager[1]

Received: 22 February 2023 / Accepted: 3 May 2024 / Published online: 23 May 2024

Abstract

We use hyperbolic wavelet regression for the fast reconstruction of high-dimensional functions having only low-dimensional variable interactions. Compactly supported periodic Chui-Wang wavelets are used for the tensorized hyperbolic wavelet basis on the torus. With a variable transformation, we are able to transform the approximation rates and fast algorithms from the torus to other domains. We perform and analyze scattered data approximation for smooth but arbitrary density functions by using a least squares method. The corresponding system matrix is sparse due to the compact support of the wavelets, which leads to a significant acceleration of the matrix vector multiplication. For non-periodic functions, we propose a new extension method. A proper choice of the extension parameter together with the piecewise polynomial Chui-Wang wavelets extends the functions appropriately. In every case, we are able to bound the approximation error with high probability. Additionally, if the function has a low effective dimension (i.e., only interactions of a few variables), we qualitatively determine the variable interactions and omit ANOVA terms with low variance in a second step in order to decrease the approximation error. This allows us to suggest an adapted model for the approximation. Numerical results show the efficiency of the proposed method.

Keywords Variable transformations · Least squares approximation · Random sampling · Wavelets · ANOVA decomposition

Mathematics Subject Classification (2010) 41A25 · 41A63 · 65D15 · 65T60

Communicated by: Ivan Oseledets

✉ Laura Weidensager
laura.weidensager@math.tu-chemnitz.de

Daniel Potts
potts@math.tu-chemnitz.de

[1] Faculty of Mathematics, Chemnitz University of Technology, 09107 Chemnitz, Germany

1 Introduction

The distribution of data points is a key component in machine learning. In many applications, a target variable has to be predicted from given high-dimensional samples. We want to reconstruct an underlying function f to give an interpretable approximation algorithm, which allows a prediction of the target variable for new samples. We consider the domains $\Omega \in \{\mathbb{T}^d, \mathbb{R}^d, [0,1]^d\}$ and also tensor products of these cases. We consider the setting of reconstructing a d-dimensional function $f\colon \Omega \to \mathbb{C}$ from discrete samples on the set of nodes $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_M\} \subset \Omega$, which are distributed to the continuous density $\varrho : \Omega \to \mathbb{R}_+$. One main aim is to also deal with an unknown density. Besides the natural question of finding a good approximation for f, we want to consider the question of interpretability, i.e., analyzing the importance of the input variables and variable interactions of the function.

Motivation
The starting point of our considerations is the question of whether it is possible to transform the good approximation results and the related fast algorithms for periodic functions on the torus $\mathbb{T}^d$ to the domain Ω. To investigate the scattered data problem on the torus, we are engaged with the sample set $\mathcal{X}$, the corresponding function values $\boldsymbol{f} = (f(\boldsymbol{x}))_{\boldsymbol{x}\in\mathcal{X}}$, and we constructed a recovery operator $S_I^{\mathcal{X}}$ in [26]. This operator computes a best least squares fit

$$S_I^{\mathcal{X}} f = \sum_{\boldsymbol{k}\in I} a_{\boldsymbol{k}} \psi_{\boldsymbol{k}}, \tag{1.1}$$

in the finite-dimensional subspace spanned by semi-orthogonal wavelets $\psi_{\boldsymbol{k}} : \mathbb{T}^d \to \mathbb{R}$ with indices in the hyperbolic cross type set I. Assuming i.i.d. uniformly samples $\mathcal{X} \subset \mathbb{T}^d$, we showed in [26, Corollary 3.22.]: Let m be the order of vanishing moments of the wavelets and the sample size have logarithmic oversampling, i.e., $|\mathcal{X}| \gtrsim r|I|\log|I|$. For $1/2 < s < m$, there is a constant $C(r,d,s) > 0$ such that for fixed Besov norm $\|f\|_{\boldsymbol{B}^s_{2,\infty}(\mathbb{T}^d)} \leqslant 1$

$$\mathbb{P}\Big(\|f - S_I^{\mathcal{X}} f\|_{L_2(\mathbb{T}^d)} \leqslant C(r,d,s)\, \frac{(\log|\mathcal{X}|)^{(d-1)(s+1/2)+s}}{|\mathcal{X}|^s} \Big) \geqslant 1 - 2|\mathcal{X}|^{-r}$$

(see Appendix 1 for the definition of the Besov space). Also, in [31], uniformly i.i.d. samples on the torus in combination with different basis functions perform well, and there are possibilities for dealing with the curse of dimensionality. The results in [26] for the periodic case serve as a basis for this paper. In many practical applications, we have to take the data set as it is and have no uniform samples available. For that reason, we study here the case where the given sample points $\mathcal{Y} \subset \Omega$ are sampled from an arbitrary (but possibly unknown) density $\varrho(\boldsymbol{y})$. In Fig. 1, we illustrate some random two-dimensional samples with respect to ϱ. We can not guarantee good approximation rates and stability if we would use these samples directly.

To this end, we investigate transformations R of the given samples. The main result of this paper is the transformed approximation operator $S_I^{\mathcal{Y}} f$. Besides the interesting results for $\Omega \in \{\mathbb{T}^d, \mathbb{R}^d\}$, we use a new extension technique for the non-periodic case $\Omega = [0, 1]^d$. Furthermore, we present a detailed error analysis.

The approach

Our main approach is to transform given samples $\mathcal{Y} \subset \Omega$ to the d-dimensional torus $\mathbb{T}^d$ by $\mathcal{X} = \mathrm{R}(\mathcal{Y})$, using the idea of inverse transform sampling: Let F be the cumulative distribution function of a distribution ϱ and $\mathcal{X} \sim \mathcal{U}([0, 1])$. Then, the random variable $F^{-1}(\mathcal{X}) \sim \varrho$ is distributed according to the distribution ϱ. Based on this, we give possibilities for constructing a transformation $\mathbb{R}$ in (3.2), which transforms the samples $\mathcal{Y} \subset \Omega^d$ to $\mathcal{X} \subset \mathbb{T}^d$ on the torus. In Fig. 1, we show an illustration of what our constructed transformation R does with the samples. In order to investigate the scattered data problem on Ω, we then use the recovery operator $S_I^{\mathcal{X}}$ from (1.1) on the torus. This operator minimizes the ℓ_2-loss function

$$\sum_{\boldsymbol{y} \in \mathcal{Y}} \left| f \circ \mathrm{R}^{-1}(\boldsymbol{y}) - S_I^{\mathcal{X}} \left(f \circ \mathrm{R}^{-1}(\boldsymbol{y}) \right) \right|^2$$

by using an iterative LSQR algorithm. To transform the approximation back, we have to apply the transformation R. We give some explicit densities and the corresponding transformations in Example 4.3 for $\Omega = \mathbb{R}^d$ and in Example 4.11 for $\Omega = [0, 1]^d$.

Our procedure coincides with transforming the function f to the function $f \circ \mathrm{R}^{-1}$, which is a function on the torus. In approximation theory, it is known that the error decay gains from the smoothness of the function. We will introduce weighted function spaces of mixed Sobolev smoothness $H^m_{\mathrm{mix}}(\Omega, \varrho)$ in Definition 4.6 and even for non-periodic functions in Definition 4.10. For the more general function class of mixed Besov spaces, we also introduce in Definition 4.9 weighted spaces $\boldsymbol{B}^s_{2,\infty}(\Omega, \varrho)$ on Ω. All our function space definitions rely on the definition of the periodic spaces. The relevant facts about Sobolev and Besov spaces of mixed smoothness on $\mathbb{T}^d$ have been collected in the Appendix 1.

Since we take the position that we can only learn the function where we have sample points, our aim is to find an approximation to the function f, which minimizes the L_2-error with respect to the density ϱ. Indeed, it shows that for functions in the defined weighted function spaces $H^m_{\mathrm{mix}}(\Omega, \varrho)$ or $\boldsymbol{B}^s_{2,\infty}(\Omega, \varrho)$, we receive in Theorem 5.1 the same approximation rates for the $L_2(\Omega, \varrho)$-error as in the periodic setting, i.e., we

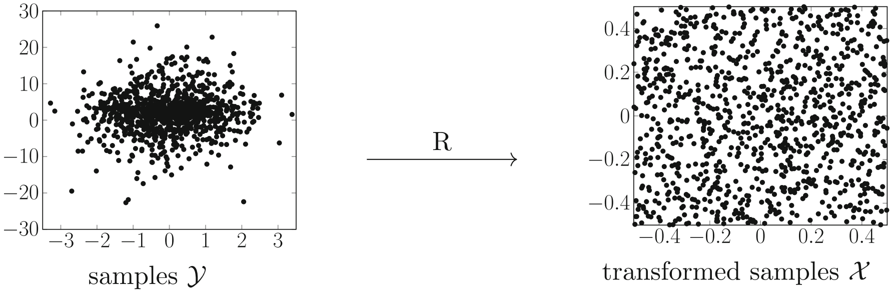


Fig. 1 Transformation of the samples in the two-dimensional case

provide that for $1/2 < s < m, r > 1$ and logarithmic oversampling $|\mathcal{Y}| \gtrsim r|I| \log |I|$, there is a constant $C(r,d,s) > 0$ such that for fixed $\|f\|_{\mathbf{B}^s_{2,\infty}(\Omega,\varrho)} \leqslant 1$

$$\mathbb{P}\Big(\|f - (S_n^{\mathcal{X}}(f \circ \mathrm{R}^{-1})) \circ \mathrm{R}\|_{L_2(\Omega,\varrho)} \leqslant C(r,d,s)\, \frac{(\log |\mathcal{Y}|)^{(d-1)(s+1/2)+s}}{|\mathcal{Y}|^s} \Big) \geqslant 1 - 2|\mathcal{Y}|^{-r}\,.$$

Wavelets have many applications in signal processing. Most commonly, they are used in compression, edge detection, noise reduction, and other signal enhancements. The broad practicality of wavelets is mainly due to the localization properties of wavelets in time and frequency, so that many signals can be sparsely represented. Hence, the hyperbolic wavelet regression is a reasonable choice for our purposes.

The approximation of non-periodic functions is more challenging than the periodic setting because of the boundary behavior. For wavelet approximation, one possibility is to construct boundary wavelets (see [11, 21]). We avoid these complicated constructions by extending the function, similar to Fourier extension [3, 5, 20]. Especially, the Chui-Wang wavelets provide an opportunity for letting the approximation extend the function itself, so that we do not have to construct the extension explicitly. We suggest to choose the transformation with a fixed but small parameter η

$$\mathrm{R}(y) = \eta + (1-\eta) \int_0^y \varrho(t)\mathrm{d}t - \tfrac{1}{2}.$$

More details are described in Sect. 5.1. Our new extension method with a properly chosen extension parameter η relies on the compact support of the Chui-Wang wavelets. With this approach, it is possible in Corollary 5.7 to end with nearly the same approximation rate as in the periodic setting.

In some applications, we usually do not know the underlying density ϱ, and we only get the samples $\mathcal{Y}$. Therefore, we first estimate the underlying density by $\mathring{\varrho}$ and construct the slightly different transformation $\mathring{\mathrm{R}}$ in Sect. 6. Using an approximation operator on $\mathbb{T}^d$ is also in this case the core idea. Naturally, the approximation error depends on the quality of the density estimation. But in Theorem 6.1, we state that we expect similar approximation results as in the case where we know the density in advance. Numerical experiments confirm this result.

For dealing with the curse of dimensionality, we introduce the *analysis of variance* (ANOVA) decomposition (see [6, 18, 27], [29, Section 3.1.6]), which decomposes the d-variate function into 2^d ANOVA terms $f_{\boldsymbol{u}}$, i.e.,

$$f(\boldsymbol{y}) = \sum_{\boldsymbol{u} \subseteq \{1,\ldots,d\}} f_{\boldsymbol{u}}(\boldsymbol{y}_{\boldsymbol{u}}).$$

Each term corresponding to $\boldsymbol{u}$ only depends on variables y_i, where $i \in \boldsymbol{u}$. The number of these variables is called *order* of the ANOVA term. However, in practical applications with high-dimensional functions, often, only the ANOVA terms of low order play a role in order to describe the function well (see [6, 12, 25, 32, 43]). For a rigorous mathematical treatment of this observation, we work with functions of low effective dimension, which allow for a truncation of the hyperbolic wavelet regression. The starting point of our work is [26], where the usage of the ANOVA decomposition was also beneficial to approximate periodic high-dimensional functions.

Mathematical modelling of complex systems often requires sensitivity analysis to determine how an output variable of interest is influenced by individual or subsets of input variables. A global sensitivity analysis constitutes the study of how the output uncertainty from a mathematical model is divvied up into distinct sources of input variation in the model. We transform the classical sensitivity analysis from the torus to a weighted function space. The transformation helps to tune the hyperbolic wavelet regression. Our main suggestion is Algorithm 1, which gives a tool for approximating high-dimensional functions from given arbitrary distributed samples from independent input variables. Furthermore, it is possible to interpret the results, since we get a knowledge about which input variables and variable interactions play a role and which do not.

One main advantage of our transformation approach is that we can deal with different domains in every variable direction. In applications, it is often the case that we have a mixture of periodic, non-periodic, and real-valued input variables on the larger domain $\mathbb{R}$. Our proposed Algorithm 1 is also applicable in these cases. Furthermore, we can use information of the densities, i.e., we handle every input variable separately, which enables a strategy to use density information where it is available. For a typical example, see Sect. 7.2.

Related work and other approaches

We will heavily use the results of [26], which gives approximation bounds and fast algorithms for the periodic setting on $\mathbb{T}^d$. In this paper, we want to generalize this to a more general (tensor product) domain Ω. Clearly, the main idea is the inverse transform sampling. But beyond that, we study function spaces on Ω, which provide enough smoothness of the transformed function to fulfill the assumptions for the periodic approximation. Further, new important aspects in this paper are the extension of non-periodic functions, similar to Fourier extension and the idea of combining density estimation and the transformation.

A nice introduction with a detailed description of the challenges in high-dimensional approximation is given in the book [1]. The change of variables was successfully used in many applications.

In [22], the authors construct a least squares approximation method for the recovery of functions from a reproducing kernel Hilbert space on $\Omega \subset \mathbb{R}^d$. The key is to construct the orthonormal basis $(\eta_k)_{k=1}^N$ in $L_2(\Omega, \varrho)$, which is in general not accessible for arbitrary or unknown densities ϱ. Also, the considerations [8, 9] are based on the knowledge of the basis η_k. With our approach, we construct the concatenated functions

$$(\eta_k(y))_{k\in I} = (\psi_k^{\text{per}}(\mathrm{R}(y)))_{k\in I},$$

which form a semi-orthogonal basis in $L_2(\Omega, \varrho)$. It is also possible to use other basis functions on $\mathbb{T}^d$ instead of the wavelet functions, but in any case, the benefit is that we have the basis in $L_2(\Omega, \varrho)$ available, even for a very general class of density functions. Furthermore, we are able to transform the fast algorithms from $\mathbb{T}^d$ to the domain Ω. A recent improvement was done in [13], where the authors used a weighted least squares algorithm with weights related to the *Christoffel function* and reduced the sampling budget by canceling the logarithmic factor. But they also assume that an orthogonal basis is known. Furthermore, in contrast to this literature, we give in Theorem 5.1 and Corollary 5.7 a concentration inequality for the approximation error based on the probabilistic Bernstein inequality in comparison to estimating the expected value.

For the examples of the Chebyshev density, which is a special case of our examples, [22, Section 10.3], [13] propose the Chebyshev polynomials $\eta_{\boldsymbol{k}}(\boldsymbol{y}) \sim \cos(\boldsymbol{k} \arccos(\boldsymbol{y}))$ as basis in $L_2([-1, 1]^d, \varrho)$ where the inner function coincides with our transformation. The case that the samples are normally distributed was considered in [33]. This approach coincides with our transformation. In Sect. 4.2, we give more details about the connection of our weighted function spaces to those in the literature. We study the case of fixed given samples $\mathcal{Y}$. In contrast to that, the task of choosing sampling points was solved successfully in [28] transforming rank-1 lattices from the torus to $\mathbb{R}^d$ or the cube $[0, 1]^d$.

Outline

This paper is organized as follows. In Sect. 2, we recall an approximation operator for periodic functions, which is based on the hyperbolic wavelet regression and the well-known ANOVA decomposition of a function on the d-dimensional torus. Section 3 describes the main idea of our approach, namely how we construct a transformation R. Section 4 is dedicated to the introduction of weighted function spaces. We study the spaces of mixed dominating Sobolev regularity in Sect. 4.1, mixed dominating Besov regularity in Sect. 4.3, and end with defining similar spaces for non-periodic functions in Sect. 4.4.

We study in this paper two settings: First, in Sect. 5, we assume that the underlying density is known. There, we show in Theorems 5.1 and Corollary 5.7 that we transfer the approximation rates from the torus to our setting on Ω. Second, we investigate in Sect. 6 the setting where we are given only the samples $\mathcal{Y}$ and no density function ϱ. Also, in this case, we are able to transfer the approximation results (see Theorem 6.1). Finally, Sect. 7 is dedicated to the presentation of Algorithm 1, which gives an interpretable high-dimensional approximation. Our theoretical results are supplemented by some numerical experiments in Sect. 7.2 that demonstrate the practical efficacy of our algorithm.

2 Preliminaries

Let us introduce the general setting and notation. Let $f : \Omega \to \mathbb{C}$ be a function on a d-dimensional domain Ω. Given are the function values $\boldsymbol{f} = (f(\boldsymbol{y}))_{\boldsymbol{y} \in \mathcal{Y}}$ at random points $\mathcal{Y} \subset \Omega$ with $|\mathcal{Y}| = M$. These samples are i.i.d. according to the density $\varrho : \Omega \to \mathbb{R}$, i.e., $\int_\Omega \varrho(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} = 1$. We will assume in this paper that $\varrho(\boldsymbol{y}) > 0$, since we otherwise omit parts of Ω where the density is equal to zero. Furthermore, we assume that the density is continuous, sufficiently smooth, and integrable. We aim to approximate the function f.

Notation

Let us introduce the weighted L_p-norm,

$$\|f\|_{L_p(\Omega,\varrho)} := \begin{cases} \left(\int_\Omega |f(\boldsymbol{y})|^p \varrho(\boldsymbol{y}) \mathrm{d}\boldsymbol{y}\right)^{1/p} & \text{if } p < \infty, \\ \sup\limits_{\boldsymbol{y}\in\Omega} |f(\boldsymbol{y})| & \text{if } p = \infty. \end{cases}$$

In the case where the density ϱ is the uniform distribution, we use the usual notations $\|f\|_{L_2(\Omega)}$ and $\|f\|_{L_\infty(\Omega)}$. We focus on the case $p = 2$, since in this case, we have the scalar product

$$\langle f, g\rangle_\varrho = \int_\Omega f(\boldsymbol{y}) g(\boldsymbol{y}) \varrho(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y}.$$

The multi-dimensional Fourier coefficients on the torus are defined by

$$c_{\boldsymbol{k}}(f) = \int_{\mathbb{T}^d} f(\boldsymbol{x}) \, \mathrm{e}^{-2\pi \mathrm{i}\langle \boldsymbol{k},\boldsymbol{x}\rangle} \, \mathrm{d}\boldsymbol{x}. \tag{2.1}$$

This allows to write every function $f \in L_2(\mathbb{T}^d)$ as a Fourier series

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{k}\in\mathbb{Z}^d} c_{\boldsymbol{k}}(f) \mathrm{e}^{2\pi \mathrm{i}\langle \boldsymbol{k},\boldsymbol{x}\rangle}.$$

In this paper, we denote by $[d]$ the set $\{1, \dots, d\}$. We work with a transformation idea, so we will always denote the d-dimensional input variable of the function f in Ω by $\boldsymbol{y}$ and the transformed values by $\boldsymbol{x} \in \mathbb{T}^d$. The subset vector is denoted by $\boldsymbol{y}_{\boldsymbol{u}} = (y_i)_{i\in\boldsymbol{u}}$ for a subset $\boldsymbol{u} \subseteq [d]$. The complement of those subsets is always with respect to $[d]$, i.e., $\boldsymbol{u}^c = [d]\backslash\boldsymbol{u}$. For an index set $\boldsymbol{u} \subseteq [d]$, we define the *order* $|\boldsymbol{u}|$ as the number of elements in $\boldsymbol{u}$.

We will study the cases where

$$\Omega = \bigtimes_{i=1}^{d} \Omega_i, \quad \Omega_i \in \{\mathbb{T}, \mathbb{R}, [0, 1]\} \text{ for all } i \in [d].$$

Note that a general interval $[a, b]$ with $b > a$ can be transferred to the unit interval via $y \mapsto \frac{y-a}{b-a}$. Similar to the vector notation, also, a subset of the domain directions is denoted by $\Omega_{\boldsymbol{u}} = \times_{i\in\boldsymbol{u}}\Omega_i$ for $\boldsymbol{u} \subseteq [d]$.

2.1 Hyperbolic wavelet regression on the torus

In this section, we introduce an approximation operator for periodic functions. For a more detailed description, see [26]. We introduce the notation $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$, for $j \in \mathbb{N}_0$, $k \in \mathbb{Z}$ and a wavelet function ψ. We use the periodization

$$\psi^{\text{per}}_{j,k}(x) = \sum_{\ell\in\mathbb{Z}} \psi_{j,k}(x + \ell),$$

where $\psi^{\mathrm{per}}_{-1,0}(x)$ denotes the scalar function as well as the tensorization

$$\psi^{\mathrm{per}}_{\boldsymbol{j},\boldsymbol{k}}(\boldsymbol{x}) = \prod_{i=1}^{d} \psi^{\mathrm{per}}_{j_i,k_i}(\boldsymbol{x}_i),$$

where $\boldsymbol{j} = (j_i)_{i=1}^d$, $j_i \in \{-1, 0, 2, \dots\}$ and $\boldsymbol{k} = (k_i)_{i=1}^d$ are multi-indices $\boldsymbol{k} \in \mathcal{I}_{\boldsymbol{j}}$. Hence, we define the sets

$$\mathcal{I}_{\boldsymbol{j}} = \bigtimes_{i=1}^{d} \begin{cases} \{0, 1, \dots 2^{j_i} - 1\} & \text{if } j_i \geqslant 0, \\ \{0\} & \text{if } j_i = -1. \end{cases}$$

Furthermore, we introduce the parameter n, which always denotes the maximal level of the used wavelets, i.e., $\mathcal{J}_n = \{\boldsymbol{j} \in \mathbb{Z}^d \mid \boldsymbol{j} \geqslant -\mathbf{1}, |\boldsymbol{j}|_1 \leqslant n\}$. For notation shortening, we introduce the index set

$$I_n = \{(\boldsymbol{j}, \boldsymbol{k}) \mid \boldsymbol{j} \in \mathcal{J}_n, \boldsymbol{k} \in \mathcal{I}_{\boldsymbol{j}}\}.$$

To construct an approximation operator which takes given samples $\mathcal{X}$, we solve the overdetermined system $\boldsymbol{A}\boldsymbol{a} = \boldsymbol{f}$, where

$$\boldsymbol{A} = (\psi^{\mathrm{per}}_{\boldsymbol{j},\boldsymbol{k}}(\boldsymbol{x}))_{\boldsymbol{x}\in\mathcal{X},(\boldsymbol{j},\boldsymbol{k})\in I_n} \in \mathbb{C}^{M\times N} \tag{2.2}$$

is the *hyperbolic wavelet matrix* with $M > N$. We will always denote the number of parameters, i.e., the number of columns of our wavelet matrix by N with $N = |I_n|$ and the number of samples by $|\mathcal{X}| = M$. A detailed connection between the maximal wavelet level n and the number of wavelet functions N can be found in [26, Lemma 3.11]. We compute the coefficients $\boldsymbol{a}$ by $\boldsymbol{a} = \left(\boldsymbol{A}^*\boldsymbol{A}\right)^{-1} \boldsymbol{A}^*\boldsymbol{f}$. We will do this iteratively by minimizing the norm $\|\boldsymbol{A}\boldsymbol{a} - \boldsymbol{f}\|_2$. This gives us the wavelet coefficients of an approximation $S_n^{\mathcal{X}} f$ to f, i.e.,

$$S_n^{\mathcal{X}} f := \sum_{\boldsymbol{j}\in\mathcal{J}_n} \sum_{\boldsymbol{k}\in\mathcal{I}_j} a_{\boldsymbol{j},\boldsymbol{k}} \psi^{\mathrm{per}}_{\boldsymbol{j},\boldsymbol{k}}. \tag{2.3}$$

A further analysis of this operator can be found in [26, Corollary 3.22]. The estimates there are valid for general wavelets, which are compactly supported, i.e.,

$$\operatorname{supp} \psi = [0, 2m - 1],$$

have vanishing moments of order m, i.e.,

$$\int_{-\infty}^{\infty} \psi(x) x^{\beta} \, \mathrm{d}x = 0, \quad \beta = 0, \dots, m - 1,$$

and the periodized wavelets form a Riesz-Basis for every index j with

$$\gamma_m \sum_{k=0}^{2^j-1} |d_{j,k}|^2 \leqslant \left\| \sum_{k=0}^{2^j-1} d_{j,k} \psi^{\mathrm{per}}_{j,k}(x) \right\|^2_{L_2(\mathbb{T})} \leqslant \delta_m \sum_{k=0}^{2^j-1} |d_{j,k}|^2. \tag{2.4}$$

Because of this semi-orthogonality, we have to use the dual basis $\psi_{\boldsymbol{j},\boldsymbol{k}}^{\mathrm{per},*}$, such that every function $f \in L_2(\mathbb{T}^d)$ can be decomposed as

$$f = \sum_{\boldsymbol{j} \geqslant -\mathbf{1}} \sum_{\boldsymbol{k} \in \mathcal{I}_{\boldsymbol{j}}} \langle f, \psi_{\boldsymbol{j},\boldsymbol{k}}^{\mathrm{per},*} \rangle \psi_{\boldsymbol{j},\boldsymbol{k}}^{\mathrm{per}}. \tag{2.5}$$

Furthermore, this decomposition gives us a connection between the wavelet coefficients,

$$\langle f, \psi_{\boldsymbol{j},\boldsymbol{k}}^{\mathrm{per}} \rangle = \sum_{\boldsymbol{k}' \in \mathcal{I}_{\boldsymbol{j}}} \langle f, \psi_{\boldsymbol{j},\boldsymbol{k}}^{\mathrm{per},*} \rangle \langle \psi_{\boldsymbol{j},\boldsymbol{k}}^{\mathrm{per}}, \psi_{\boldsymbol{j},\boldsymbol{k}'}^{\mathrm{per}} \rangle.$$

In [26], we also introduced and analyzed the projection operator onto the wavelet space,

$$P_n f := \sum_{\boldsymbol{j} \in \mathcal{J}_n} \sum_{\boldsymbol{k} \in \mathcal{I}_{\boldsymbol{j}}} \langle f, \psi_{\boldsymbol{j},\boldsymbol{k}}^{\mathrm{per}*} \rangle \psi_{\boldsymbol{j},\boldsymbol{k}}^{\mathrm{per}}. \tag{2.6}$$

The following estimates are a short summary of the results from [26] for periodic functions. For the definition of the function spaces, see Appendix 1.

$$\|f - P_n f\|_{L_2(\mathbb{T}^d)} \lesssim 2^{-mn} n^{(d-1)/2} \, \|f\|_{H^m_{\mathrm{mix}}(\mathbb{T}^d)} \tag{2.7}$$

$$\|f - P_n f\|_{L_\infty(\mathbb{T}^d)} \lesssim 2^{-n(m-1/2)} n^{(d-1)} \, \|f\|_{H^m_{\mathrm{mix}}(\mathbb{T}^d)}, \tag{2.8}$$

$$\mathbb{P}\left(\|f - S_n^{\mathcal{X}} f\|^2_{L_2(\mathbb{T}^d)} \lesssim 2^{-2nm} n^{d-1} \, \|f\|^2_{H^m_{\mathrm{mix}}(\mathbb{T}^d)} \right) \geqslant 1 - 2\, M^{-r}, \tag{2.9}$$

$$\mathbb{P}\left(\|f - S_n^{\mathcal{X}} f\|^2_{L_2(\mathbb{T}^d)} \lesssim 2^{-2ns} n^{d-1} \, \|f\|^2_{\boldsymbol{B}^s_{2,\infty}(\mathbb{T}^d)} \right) \geqslant 1 - 2\, M^{-r}, \quad \tfrac{1}{2} < s < m \tag{2.10}$$

where the last result holds for some $r > 1$ if we have $M \gtrsim rN \log N$ and uniformly i.i.d. samples $\mathcal{X}$. We will focus our numerical experiments on Chui-Wang-wavelets, where we will always denote by m the *order of the wavelets*, which denotes the number of vanishing moments of the wavelets.

2.2 The ANOVA decomposition

The curse of dimensionality comes into play whenever one deals with high-dimensional functions. The aim of sensitivity analysis is to describe the structure of multivariate periodic functions f and to analyze the influence of each variable. A frequently used concept is the following [6, 18, 27].

Definition 2.1 Let f be in $L_2(\mathbb{T}^d)$. For a subset $\boldsymbol{u} \subseteq [d]$, we define the *ANOVA (analysis of variance) terms* by

$$f_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}}) = \int_{\mathbb{T}^{d-|\boldsymbol{u}|}} f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}_{\boldsymbol{u}^c} - \sum_{\boldsymbol{v} \subset \boldsymbol{u}} f_{\boldsymbol{v}}(\boldsymbol{x}_{\boldsymbol{v}}). \tag{2.11}$$

The *ANOVA decomposition* of a function $f\colon \mathbb{T}^d \to \mathbb{C}$ is then given by

$$f(\boldsymbol{x}) = f_\varnothing + \sum_{i=1}^{d} f_{\{i\}}(x_i) + \sum_{i \neq j=1}^{d} f_{\{i,j\}}(x_i, x_j) + \cdots + f_{[d]}(\boldsymbol{x}) = \sum_{\boldsymbol{u} \subseteq [d]} f_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}}). \tag{2.12}$$

The terms (2.11) are the unique decomposition (2.12), such that they have mean zero and are pairwise orthogonal. Additionally, the decomposition (2.5) of a function $f \in L_2(\mathbb{T}^d)$ in terms of wavelets can be written in ANOVA terms, i.e.,

$$f_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}}) = \sum_{(\boldsymbol{j},\boldsymbol{k}) \in I^{\boldsymbol{u}}} a_{\boldsymbol{j},\boldsymbol{k}} \psi^{\mathrm{per}}_{\boldsymbol{j},\boldsymbol{k}}(\boldsymbol{x}), \quad I^{\boldsymbol{u}} := \{(\boldsymbol{j},\boldsymbol{k}) \mid \boldsymbol{j}_{\boldsymbol{u}^c} = -\boldsymbol{1}, \boldsymbol{k} \in \mathcal{I}_{\boldsymbol{j}}\}.$$

The same connection holds true for a truncated wavelet decomposition with $|\boldsymbol{j}|_1 \leqslant n$. That means, all hyperbolic indices $(\boldsymbol{j},\boldsymbol{k}) \in I_n$ can be decomposed in a disjoint union of index sets belonging to one ANOVA term with index $\boldsymbol{u} \subset [d]$, i.e.,

$$I_n = \dot{\bigcup_{\boldsymbol{u} \subseteq [d]}} I_n^{\boldsymbol{u}}, \quad I_n^{\boldsymbol{u}} = \{(\boldsymbol{j},\boldsymbol{k}) \in I_n \mid \boldsymbol{j}_{\boldsymbol{u}^c} = -\boldsymbol{1}\}. \tag{2.13}$$

This connection is illustrated in Fig. 2 for $d = 3$ and $n = 3$. The crucial property is that for an index $\boldsymbol{u}$, the corresponding functions $\psi^{\mathrm{per}}_{\boldsymbol{j},\boldsymbol{k}}$ have to be constant in all directions $i \notin \boldsymbol{u}$, i.e., $\boldsymbol{j}_{\boldsymbol{u}^c} = -\boldsymbol{1}$. For further details, see [26, Section 4].

To get a notion of the importance of single terms compared to the entire function, we define the *variance* of a function by

$$\sigma^2(f) := \int_{\mathbb{T}^d} \left| f(\boldsymbol{x}) - \int_{\mathbb{T}^d} f(\boldsymbol{x}')\,\mathrm{d}\boldsymbol{x}' \right|^2 \mathrm{d}\boldsymbol{x} = \int_{\mathbb{T}^d} |f(\boldsymbol{x})|^2\,\mathrm{d}\boldsymbol{x} - f_\varnothing^2.$$

The idea of the ANOVA decomposition is to analyze which combinations of the input variables x_j play a role for the approximation of f. The variances of the ANOVA terms indicate their importance; hence, we do the following. For subsets $\boldsymbol{u} \subseteq [d]$ with $\boldsymbol{u} \neq \varnothing$, the *global sensitivity indices* (GSI) [39] are defined as

$$S(\boldsymbol{u}, f) := \frac{\sigma^2(f_{\boldsymbol{u}})}{\sigma^2(f)} \in [0, 1], \tag{2.14}$$

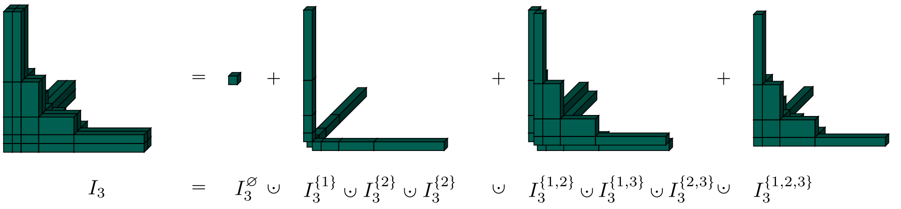

=
+
+
+
$I_3 = I_3^{\varnothing} \dot\cup I_3^{\{1\}} \dot\cup I_3^{\{2\}} \dot\cup I_3^{\{2\}} \dot\cup I_3^{\{1,2\}} \dot\cup I_3^{\{1,3\}} \dot\cup I_3^{\{2,3\}} \dot\cup I_3^{\{1,2,3\}}$

Fig. 2 Illustration of the ANOVA indices of a three-dimensional function. Every cuboid belongs to one index $\boldsymbol{j}$. The size represents the number of translation indices $\boldsymbol{k} \in \mathcal{I}_{\boldsymbol{j}}$, which gives this dyadic structure. All indices in I_3 are decomposed into indices belonging to the ANOVA terms with index $\boldsymbol{u} \subseteq [3]$

where the variance of the ANOVA term $f_{\boldsymbol{u}}$ is

$$\sigma^2(f_{\boldsymbol{u}}) = \int_{\mathbb{T}^{|\boldsymbol{u}|}} |f_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}})|^2 \,\mathrm{d}\boldsymbol{x}_{\boldsymbol{u}},$$

since the mean of the ANOVA terms is zero. The $L_2(\mathbb{T}^d)$-orthogonality of the ANOVA terms implies that the variance of $f(\boldsymbol{x})$ for $L_2(\mathbb{T}^d)$-functions f can be decomposed as

$$\sigma^2(f) = \sum_{\substack{\boldsymbol{u}\subseteq[d]\\ \boldsymbol{u}\neq\varnothing}} \sigma^2(f_{\boldsymbol{u}}).$$

This implies

$$\sum_{\substack{\boldsymbol{u}\subseteq[d]\\ \boldsymbol{u}\neq\varnothing}} S(\boldsymbol{u}, f) = 1.$$

The global sensitivity index $S(\boldsymbol{u}, f)$ represents the proportion of the variance of $f(\boldsymbol{x})$ explained by the interaction between the variables indexed by $\boldsymbol{u}$. These indices can also be computed using only the wavelet coefficients of a function with the connection (2.13).

3 Transformations of functions to the torus

In this chapter, we introduce our main approach: the transformation procedure. We transform a function defined on some domain Ω to the torus, use the well-studied approximation operator for periodic functions, and transform the result back to a function defined on Ω. **The univariate setting**

The basis for our transformation is the *cumulative distribution function* $F : \Omega \to [0, 1]$, which fulfills

$$\frac{\mathrm{d}}{\mathrm{d}y} F(y) = \varrho(y). \tag{3.1}$$

Subsequently, we identify $[0, 1]$ with the torus by the bijective mapping $y \mapsto y - \frac{1}{2}$. Similarly to the cumulative distribution function, we define the *transformation*

$$\mathrm{R}\colon \Omega \to [-\tfrac{1}{2}, \tfrac{1}{2}], \quad \mathrm{R}(y) := \begin{cases} \displaystyle\int_{-1/2}^{y} \varrho(t)\,\mathrm{d}t - \tfrac{1}{2} & \text{if } \Omega = \mathbb{T}, \\ \displaystyle\int_{-\infty}^{y} \varrho(t)\,\mathrm{d}t - \tfrac{1}{2} & \text{if } \Omega = \mathbb{R}, \\ \displaystyle\eta + (1-\eta)\int_{0}^{y} \varrho(t)\,\mathrm{d}t - \tfrac{1}{2}, \quad 0 \leqslant \eta \ll 1 & \text{if } \Omega = [0, 1]. \end{cases} \tag{3.2}$$

This transformation R gives us a possibility to transfer the function f to a function $f \circ \mathrm{R}^{-1}$, which has its domain on the torus. Since we require that the density ϱ is positive, the cumulative distribution function is strictly monotone increasing and has a well-defined inverse function R^{-1}. In the case of a non-periodic function, we have to use an extension with parameter η, since otherwise the transformed function is not even continuous. Our transformation R transforms the function f from $[0, 1]$ to

$[-\nicefrac{1}{2}+\eta, \nicefrac{1}{2}]$, which means we extend the function on the boundary $[-\nicefrac{1}{2}, -\nicefrac{1}{2}+\eta]$ to receive a periodic function $\tilde{f}\colon \mathbb{T} \to \mathbb{C}$ with

$$\tilde{f}\big|_{[-\nicefrac{1}{2}+\eta,\nicefrac{1}{2}]} = f \circ \mathrm{R}^{-1}.$$

We give more details about this extension in Sect. 5.1. The connection between the functions is illustrated here:

$$\begin{tikzcd} \Omega \arrow[rr, shift left, "\mathrm{R}"] \arrow[dr, "f"'] & & \mathbb{T} \arrow[ll, shift left, "\mathrm{R}^{-1}"] \arrow[dl, "f\circ\mathrm{R}^{-1} \text{ or } \tilde{f}"] \\ & \mathbb{C} & \end{tikzcd}$$

The transformation R has the property that

$$\frac{\mathrm{d}}{\mathrm{d}y}\mathrm{R}(y) = \begin{cases} \varrho(y) & \text{if } \Omega \in \{\mathbb{T}, \mathbb{R}\}, \\ (1-\eta)\,\varrho(y) & \text{if } \Omega = [0,1]. \end{cases}$$

A variable substitution shows for $\Omega \in \{\mathbb{T}, \mathbb{R}\}$ the important relation

$$\begin{aligned} \|f\|^2_{L_2(\Omega,\varrho)} &= \int_\Omega |f(y)|^2 \varrho(y)\,\mathrm{d}y = \int_\mathbb{T} |f(\mathrm{R}^{-1}(x))|^2 \varrho(\mathrm{R}^{-1}(x))(\mathrm{R}^{-1})'(x)\,\mathrm{d}x \\ &= \int_\mathbb{T} |f(\mathrm{R}^{-1}(x))|^2 \varrho(\mathrm{R}^{-1}(x)) \tfrac{1}{\mathrm{R}'(\mathrm{R}^{-1}(x))}\,\mathrm{d}x = \int_\mathbb{T} |f(\mathrm{R}^{-1}(x))|^2\,\mathrm{d}x = \|f\circ\mathrm{R}^{-1}\|^2_{L_2(\mathbb{T})}. \end{aligned} \tag{3.3}$$

In the case where $\Omega = [0,1]$, we have

$$\begin{aligned} \|f\|^2_{L_2([0,1],\varrho)} = \int_0^1 |f(y)|^2 \varrho(y)\,\mathrm{d}y &= \frac{1}{(1-\eta)} \int_{-\nicefrac{1}{2}+\eta}^{\nicefrac{1}{2}} |f(\mathrm{R}^{-1}(x))|^2\,\mathrm{d}\boldsymbol{x} \\ &= \frac{1}{(1-\eta)}\,\|f\circ\mathrm{R}^{-1}\|^2_{L_2([-\nicefrac{1}{2}+\eta,\nicefrac{1}{2}])}. \end{aligned}$$

This relation between the L_2-norms motivates to transform the samples $\mathcal{Y} \subset \Omega$ to the transformed samples $\mathcal{X} = \mathrm{R}(\mathcal{Y})$ and then use an approximation operator on $\mathbb{T}$. In this paper, we use the operator $S_n^\mathcal{X}$, defined in (2.3). But in general, the transformation can be applied to any approximation operator on $\mathbb{T}$. At the end, we receive the approximation

$$(S_n^\mathcal{X}(f\circ\mathrm{R}^{-1}))\circ\mathrm{R}, \tag{3.4}$$

which takes the given sample points $\mathcal{Y}$ and gives back a function defined on Ω.

In fact, we change the function and approximate the transformed function $f\circ\mathrm{R}^{-1}$, which is a function on $\mathbb{T}$. For the approximation operator on $\mathbb{T}$, it is known that the smoother the function, the better the approximation. Indeed, it is not clear whether the transformed function $f\circ\mathrm{R}^{-1}$ inherits the smoothness of the function f itself, if the density is smooth enough. So far, we do not know which regularity the transformed function $f\circ\mathrm{R}^{-1}$ has. In the following, we will show that if we request more regularity from the density ϱ, the regularity in the Sobolev and Besov norm is preserved under the transformation, i.e., our aim is to introduce suitable weighted norms on Ω, namely $\|f\|_{H^s(\Omega,\varrho)}$ and $\|f\|_{\boldsymbol{B}^s_{2,\infty}(\Omega,\varrho)}$, such that

$$\|f \circ \mathrm{R}^{-1}\|_{H^s(\mathbb{T})} \leqslant \|f\|_{H^s(\Omega,\varrho)} \tag{3.5}$$

$$\|f \circ \mathrm{R}^{-1}\|_{\boldsymbol{B}^s_{2,\infty}(\mathbb{T})} \leqslant \|f\|_{\boldsymbol{B}^s_{2,\infty}(\Omega,\varrho)} \,. \tag{3.6}$$

The multivariate setting

In the multivariate setting, we consider the domain $\Omega = \times_{i=1}^d \Omega_i$ with $\Omega_i \in \{\mathbb{T}, \mathbb{R}, [0,1]\}$ for $i \in [d]$. We require that the input variables y_i are independent, which means that the density $\varrho(\boldsymbol{y})$ is a product measure,

$$\varrho(\boldsymbol{y}) = \prod_{i=1}^{d} \varrho_i(y_i). \tag{3.7}$$

We build up a d-dimensional transformation $\mathrm{R}\colon \Omega \to \mathbb{T}^d$ from one-dimensional transformations (3.2) by

$$\mathrm{R}(\boldsymbol{y}) := (\mathrm{R}_1(y_1), \ldots, \mathrm{R}_d(y_d)) \quad \text{with} \quad \frac{\mathrm{d}}{\mathrm{d}y_i}\mathrm{R}_i(y_i) = \begin{cases} \varrho_i(y_i) & \text{if } \Omega_i \in \{\mathbb{T}, \mathbb{R}\}, \\ (1-\eta)\varrho_i(y_i) & \text{if } \Omega_i = [0,1]. \end{cases} \tag{3.8}$$

From time to time, we use the notation $\mathrm{R}_{\boldsymbol{u}}(\boldsymbol{y}_{\boldsymbol{u}}) = (\mathrm{R}_i(y_i))_{i\in\boldsymbol{u}}$, which is similar to the notation for vectors with index $\boldsymbol{u}$. The inverse transformation is given by

$$\mathrm{R}^{-1}(\boldsymbol{x}) = \left(\mathrm{R}_1^{-1}(x_1), \ldots, \mathrm{R}_d^{-1}(x_d)\right) \quad \text{with} \quad \frac{\mathrm{d}}{\mathrm{d}x_j}\mathrm{R}_i^{-1}(x_i) = \delta_{i,j} \begin{cases} \frac{1}{\varrho_i(\mathrm{R}_i^{-1}(x_i))} & \text{if } \Omega_i \in \{\mathbb{T}, \mathbb{R}\}, \\ \frac{1}{(1-\eta)\,\varrho_i(\mathrm{R}_i^{-1}(x_i))} & \text{if } \Omega_i = [0,1]. \end{cases} \tag{3.9}$$

The relation that we will use through this paper can be seen in this illustration:

samples:

$$\begin{tikzcd} \boldsymbol{y} \in \mathcal{Y} \arrow[r, shift left, "\mathrm{R}"] & \boldsymbol{x} \in \mathcal{X} \arrow[l, shift left, "\mathrm{R}^{-1}"] \end{tikzcd}$$

domain:

$$\begin{tikzcd} \Omega \arrow[rr, shift left, "\mathrm{R}"] \arrow[dr, "f"'] & & \mathbb{T}^d \arrow[ll, shift left, "\mathrm{R}^{-1}"] \arrow[dl, "f\circ\mathrm{R}^{-1} \text{ or } \tilde{f}"] \\ & \mathbb{C} & \end{tikzcd}$$

By the observation that the Jacobi matrix

$$(\mathrm{D}(\mathrm{R}^{-1})(\boldsymbol{x}))_{i,j} := \frac{\partial}{\partial x_j}\mathrm{R}_i^{-1}(\boldsymbol{x}) = \delta_{i,j} \begin{cases} \frac{1}{\varrho_i(\mathrm{R}_i^{-1}(x_i))} & \text{if } \Omega_i \in \{\mathbb{T}, \mathbb{R}\}, \\ \frac{1}{(1-\eta)\,\varrho_i(\mathrm{R}_i^{-1}(x_i))} & \text{if } \Omega_i = [0,1]. \end{cases}$$

is a diagonal matrix because of the product structure (3.8) of our transformation, it follows that, similar as in the univariate case (3.3),

$$\begin{aligned}\|f\|^2_{L_2(\Omega,\varrho)} &= \int_\Omega |f(\boldsymbol{y})|^2 \varrho(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} = \int_{\mathbb{T}^d} |f(\mathrm{R}^{-1}(\boldsymbol{x}))|^2 \varrho(\mathrm{R}^{-1}(\boldsymbol{x}))|\det\left(\mathrm{D}(\mathrm{R}^{-1})(\boldsymbol{x})\right)|\,\mathrm{d}\boldsymbol{x}\\ &= \int_{\mathbb{T}^d} |f(\mathrm{R}^{-1}(\boldsymbol{x}))|^2\,\mathrm{d}\boldsymbol{x} = \frac{1}{(1-\eta)^{\tilde{d}}}\,\|f\circ\mathrm{R}^{-1}\|^2_{L_2(\mathbb{T}^d)}\,,\end{aligned} \tag{3.10}$$

where $\tilde{d} = |\{i \in [d] \mid \Omega_i = [0,1]\}|$ is the dimension of the non-periodic variables of f. The norm equality (3.10) ensures that the transformation R preserves the L_2-norm of the function f, up to some factor for the non-periodic setting.

One main advantage of our transformation approach is that we have in addition a semi-orthogonal system on Ω with respect to ϱ, i.e.,

$$\langle \psi^{\mathrm{per}}_{\boldsymbol{j},\boldsymbol{k}} \circ \mathrm{R}^{-1}, \psi^{\mathrm{per}}_{\boldsymbol{j}',\boldsymbol{k}'} \circ \mathrm{R}^{-1}\rangle_\varrho = \delta_{\boldsymbol{j},\boldsymbol{j}'}\, \langle \psi^{\mathrm{per}}_{\boldsymbol{j},\boldsymbol{k}} \psi^{\mathrm{per}}_{\boldsymbol{j},\boldsymbol{k}'}\rangle.$$

The next chapter is dedicated to the introduction of weighted function spaces on Ω, such that also the smoothness of the function is inherited, i.e., we want to generalize the equations (3.5) and (3.6) to the function spaces of dominating mixed derivatives.

3.1 The transformation R meets the ANOVA decomposition

For an L_2-function on the domain $\Omega = \times_{i=1}^d \Omega_i$ with $\Omega_i \in \{\mathbb{T}, \mathbb{R}, [0,1]\}$ for $i = 1, \dots, d$ with respect to the density ϱ, it is possible to define a generalized ANOVA decomposition. We assume in this paper that the input variables y_i are independent, which means that $\varrho(\boldsymbol{y})$ has a product structure (3.7). Hence, for an ANOVA index $\varnothing \neq \boldsymbol{u} \subset \{1, \dots, d\}$, we define the marginal distributions

$$\varrho_{\boldsymbol{u}} \colon \Omega_{\boldsymbol{u}} \to \mathbb{R}, \quad \varrho_{\boldsymbol{u}}(\boldsymbol{y}_{\boldsymbol{u}}) := \prod_{i\in\boldsymbol{u}} \varrho_i(y_i).$$

Then, the ANOVA decomposition with respect to the measure ϱ is defined by

$$f(\boldsymbol{y}) = \sum_{\boldsymbol{u}\subseteq[d]} f_{\boldsymbol{u}}(\boldsymbol{y}_{\boldsymbol{u}}), \tag{3.11}$$

where the ANOVA terms are expressed, analogously to (2.11) by a recursive formula

$$\begin{aligned} f_\varnothing &= \int_\Omega f(\boldsymbol{y})\varrho(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}\\ f_{\boldsymbol{u}}(\boldsymbol{y}_{\boldsymbol{u}}) &= \int_{\Omega_{\boldsymbol{u}^c}} f(\boldsymbol{y})\varrho_{\boldsymbol{u}^c}(\boldsymbol{y}_{\boldsymbol{u}^c})\,\mathrm{d}\boldsymbol{y}_{\boldsymbol{u}^c} - \sum_{\boldsymbol{v}\subset\boldsymbol{u}} f_{\boldsymbol{v}}(\boldsymbol{y}_{\boldsymbol{v}}),\end{aligned} \tag{3.12}$$

see also [14, 25, 34] for the case in $\mathbb{R}^d$. Our main idea is to transform a function f from Ω to the torus $\mathbb{T}^d$. Using Definition 2.1, we have a decomposition of periodic functions on the torus, i.e., for the function $f \circ \mathrm{R}^{-1}$. If we transform this decomposition back to Ω, we receive the decomposition (3.11). This can be seen by the following.

Lemma 3.1 *Let $\Omega_i \in \{\mathbb{T}, \mathbb{R}\}$ for all $i \in [d]$. The ANOVA terms defined in* (3.12) *are the same as the transformed terms of the periodic function $f \circ \mathrm{R}^{-1}$ with the transformation defined in* (3.2)*, i.e.,*

$$(f \circ \mathrm{R}^{-1})_{\boldsymbol{u}}(\mathrm{R}_{\boldsymbol{u}}(\boldsymbol{y}_{\boldsymbol{u}})) = f_{\boldsymbol{u}}(\boldsymbol{y}_{\boldsymbol{u}}).$$

Proof We defined the ANOVA terms on Ω as well as the terms on $\mathbb{T}^d$ recursively (see (2.11) and (3.11)). Hence, we show by induction over the order $|\boldsymbol{u}|$ with the help of the substitution $\mathrm{R}(\boldsymbol{y}) = \boldsymbol{x}$:

$$\begin{aligned}
(f \circ \mathrm{R}^{-1})_{\varnothing} &= \int_{\mathbb{T}^d} f(\mathrm{R}^{-1}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} = \int_{\Omega} f(\boldsymbol{y})\varrho(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} = f_{\varnothing},\\
(f \circ \mathrm{R}^{-1})_{\boldsymbol{u}}(\mathrm{R}_{\boldsymbol{u}}(\boldsymbol{y}_{\boldsymbol{u}})) = (f \circ \mathrm{R}^{-1})_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}}) &= \int_{\mathbb{T}^{|\boldsymbol{u}^c|}} f(\mathrm{R}^{-1}(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x}_{\boldsymbol{u}^c} - \sum_{\boldsymbol{v} \subset \boldsymbol{u}} (f \circ \mathrm{R}^{-1})_{\boldsymbol{v}}(\boldsymbol{x}_{\boldsymbol{v}})\\
&= \int_{\Omega_{\boldsymbol{u}^c}} f(\boldsymbol{y})\varrho_{\boldsymbol{u}^c}(\boldsymbol{y}_{\boldsymbol{u}^c})\,\mathrm{d}\boldsymbol{y}_{\boldsymbol{u}^c} - \sum_{\boldsymbol{v} \subset \boldsymbol{u}} (f \circ \mathrm{R}^{-1})_{\boldsymbol{v}}(\mathrm{R}_{\boldsymbol{u}}(\boldsymbol{y}_{\boldsymbol{v}}))\\
&= \int_{\Omega_{\boldsymbol{u}^c}} f(\boldsymbol{y})\varrho_{\boldsymbol{u}^c}(\boldsymbol{y}_{\boldsymbol{u}^c})\,\mathrm{d}\boldsymbol{y}_{\boldsymbol{u}^c} - \sum_{\boldsymbol{v} \subset \boldsymbol{u}} f_{\boldsymbol{v}}(\boldsymbol{y}_{\boldsymbol{v}}) = f_{\boldsymbol{u}}(\boldsymbol{y}_{\boldsymbol{u}}).
\end{aligned}$$

This gives the assertion. □

The decomposition (3.11) of $f \in L_2(\Omega, \varrho)$ preserves the orthogonality of the ANOVA terms, since a simple substitution similar to (3.10) shows that

$$\begin{aligned}
\langle f_{\boldsymbol{u}}, f_{\boldsymbol{v}}\rangle_{\varrho} &= \int_{\Omega} f_{\boldsymbol{u}}(y_{\boldsymbol{u}}) f_{\boldsymbol{v}}(\boldsymbol{y}_{\boldsymbol{v}})\varrho(\boldsymbol{y})\mathrm{d}\boldsymbol{y} = \int_{\mathbb{T}^d} (f \circ \mathrm{R}^{-1})_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}})(f \circ \mathrm{R}^{-1})_{\boldsymbol{v}}(\boldsymbol{x}_{\boldsymbol{u}})\mathrm{d}\boldsymbol{x}\\
&= \begin{cases} 0 & \text{if } \boldsymbol{v} \neq \boldsymbol{u},\\ \|(f \circ \mathrm{R}^{-1})_{\boldsymbol{u}}\|^2_{L_2(\mathbb{T}^d)} & \text{if } \boldsymbol{v} = \boldsymbol{u}. \end{cases}
\end{aligned}$$

Hence, the variance of the ANOVA term of the transformed function $(f \circ \mathrm{R}^{-1})_{\boldsymbol{u}}$ is equal to the variance of $f_{\boldsymbol{u}}$ with respect to the density ϱ,

$$\sigma_{\varrho}^2(f_{\boldsymbol{u}}) := \int_{\Omega_{\boldsymbol{u}}} |f_{\boldsymbol{u}}(\boldsymbol{y}_{\boldsymbol{u}})|^2 \prod_{i \in \boldsymbol{u}} \varrho_i(y_i)\,\mathrm{d}\boldsymbol{y}_{\boldsymbol{u}}. \tag{3.13}$$

Analogously to the unweighted case (2.14), we define the *global sensitivity indices* for functions defined on Ω by

$$S(\boldsymbol{u}, f) := \frac{\sigma_{\varrho}^2(f_{\boldsymbol{u}})}{\sigma_{\varrho}^2(f)} \in [0, 1]. \tag{3.14}$$

Remark 3.2 (Dependent input variables) Note that there is no natural way to decompose f into ANOVA terms for dependent input variables. Consider the extremal case where $y_i = y_j$ for $i \neq j$: It is not possible to say which proportion of the variance belongs to f_i or f_j. Problems arise when the input variables are correlated. If we integrate over some distribution, when in reality features are dependent, we create a new data set that deviates from the joint distribution and extrapolates to unlikely combinations of features, which can indicate unwanted variances for feature decompositions. Thus, there has to be a precomputation step to avoid such dependencies. It would be possible to preprocess the given data by a PCA and a linear data transformation. Furthermore, there are approaches to generalize the ANOVA decomposition to dependent variables (see, for example, [19, 35]). The generalized ANOVA decomposition is very difficult to estimate, and the generalization of our approach to this setting is behind the scope of our paper and provides an opportunity for further research.

4 Weighted function spaces

In this chapter, we introduce weighted function spaces on Ω, which generalize smoothness from periodic functions to functions defined on Ω using the transformation R. The general idea is to study the smoothness of the concatenated function $f \circ \mathrm{R}^{-1}$, since in the periodic setting we know from [26] the higher the smoothness, the better approximation results using hyperbolic wavelet regression. On the torus, there are results for the Sobolev and Besov regularity, which is based on the wavelet coefficient decay in these spaces. For that reason, we study these two cases, in which functions on Ω are transformed to a smooth function on $\mathbb{T}^d$. In Sect. 4.1, we study the Sobolev norm (A.1) or (A.2), which requests the norm inequality (3.5) to preserve smoothness. For that reason, we introduce in Definition 4.6 a weighted Sobolev norm on Ω, which can be generalized to fractional smoothness by (4.10). In Sect. 4.2, we present the relation to function spaces already known from the literature. To preserve Besov regularity with the transformation R, we use in Sect. 4.3 the characterization (A.3) and we introduce the weighted Besov spaces in Definition 4.9, which fulfill (3.6).
First, we will study in the following the case where $\Omega_i \in \{\mathbb{T}, \mathbb{R}\}$ for all $i \in [d]$. In Sect. 4.4, we then show that the non-periodic setting is similar, up to some slight modifications.

4.1 Weighted Sobolev spaces

For measuring the smoothness of the transformed function $f \circ \mathrm{R}^{-1}$, we have to calculate the derivatives of the concatenation $f \circ \mathrm{R}^{-1}$ (see Definition (A.1)). We use the transformation (3.2) and consider in this subsection only the case $\Omega_i \in \{\mathbb{T}, \mathbb{R}\}$. The slight modification for non-periodic functions, i.e., $\eta > 0$ is described in Sect. 4.4.

The univariate setting
We use Faá di Bruno's formula, which generalizes the chain rule for $\alpha \geqslant 1$ to

$$\frac{\mathrm{d}^\alpha}{\mathrm{d}x^\alpha} f(\mathrm{R}^{-1}(x)) = \sum_{i=1}^{\alpha} f^{(i)}(y)\, B_{\alpha,i}((\mathrm{R}^{-1})^{(1)}(x), (\mathrm{R}^{-1})^{(2)}(x), \dots, (\mathrm{R}^{-1})^{(\alpha-i+1)}(x)),$$

where $B_{\alpha,i}$ are the Bell polynomials

$$B_{\alpha,i}(x_1,\ldots,x_{\alpha-i+1}) = \sum \frac{\alpha!}{j_1! j_2 \cdots j_{\alpha-i+1}!} \prod_{k=1}^{\alpha-i+1} \left(\frac{x_k}{k!}\right)^{j_k},$$

where the sum is taken over all sequences $j_1, j_2, \ldots, j_{\alpha-i+1}$ of non-negative integers, such that these two conditions are satisfied:

$$\sum_{k=1}^{\alpha-i+1} j_k = i, \quad \sum_{k=1}^{\alpha-i+1} k \cdot j_k = \alpha.$$

We use the substitution $\mathrm{R}(x) = y$. Note that for the differentials of the inverse transformation, R^{-1} holds

$$(\mathrm{R}^{-1})^{(1)}(x) = \frac{1}{\varrho(y)}, \quad (\mathrm{R}^{-1})^{(2)}(x) = \frac{-\varrho^{(1)}(y)}{\varrho^3(y)}, \quad (\mathrm{R}^{-1})^{(3)}(x) = \frac{\varrho^{(2)}(y)}{\varrho^4(y)} + \frac{3\,(\varrho^{(1)}(y))^2}{\varrho^5(y)},$$
$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{1}{\varrho(y)}\right)^k = k\left(\frac{1}{\varrho(y)}\right)^{k-1}\left(-\frac{1}{\varrho^2(y)}\right)\frac{\varrho^{(1)}(y)}{\varrho(y)} = -k\left(\frac{1}{\varrho(y)}\right)^{k+2}\varrho^{(1)}(y).$$

Thus, every derivative of $(\mathrm{R}^{-1})^{(k)}(x)$ can be expressed by a term containing only derivatives of $\varrho(y)$ up to order $k-1$ as well as power of $\frac{1}{\varrho(y)}$ up to order $2k-1$. This allows us to shorten the notation by

$$B_{\alpha,i}(y) := B_{\alpha,i}((\mathrm{R}^{-1})^{(1)}(x), (\mathrm{R}^{-1})^{(2)}(x), \ldots, (\mathrm{R}^{-1})^{(\alpha-i+1)}(x)). \tag{4.1}$$

With this notation, Faá di Bruno's formula reads as

$$\frac{\mathrm{d}^\alpha}{\mathrm{d}x^\alpha} f(\mathrm{R}^{-1}(x)) = \sum_{i=1}^{\alpha} B_{\alpha,i}(y)\, f^{(i)}(y).$$

The Bell polynomial $B_{\alpha,i}$ can be expressed in terms of derivatives of $\varrho(y)$ up to order $\alpha-1$ and powers of ϱ up to order $4\alpha-2i-1$. We have for small indices,

$$\begin{aligned}
&B_{1,1}(y) = \tfrac{1}{\varrho(y)}, \\
&B_{2,1}(y) = -\tfrac{\varrho'(y)}{\varrho^3(y)}, && B_{2,2}(y) = \tfrac{1}{\varrho^2(y)}, \\
&B_{3,1}(y) = -\tfrac{\varrho^{(2)}(y)}{\varrho^4(y)} + \tfrac{3(\varrho'(y))^2}{\varrho^5(y)}, && B_{3,2}(y) = \tfrac{-3\varrho'(y)}{\varrho^4(y)}, && B_{3,3}(y) = \tfrac{1}{\varrho^3(y)}.
\end{aligned} \tag{4.2}$$

The $L_2(\mathbb{T})$-norm of the derivatives of $f \circ \mathrm{R}^{-1}$ can thus be expressed as

$$\begin{aligned}
\left\|\frac{\mathrm{d}^\alpha}{\mathrm{d}x^\alpha} f \circ \mathrm{R}^{-1}(x)\right\|_{L_2(\mathbb{T})}^2 &= \int_{\mathbb{T}} \left|\sum_{k=1}^{\alpha} B_{\alpha,k}(y)\,\mathrm{D}^k f(y)\right|^2 \mathrm{d}x \leqslant \sum_{k=1}^{\alpha} \int_{\Omega} \left|B_{\alpha,k}(y)\,\mathrm{D}^k f(y)\right|^2 \varrho(y)\,\mathrm{d}y \\
&= \sum_{k=1}^{\alpha} \left\|\mathrm{D}^k f(y)\right\|_{L_2(\Omega, |B_{\alpha,k}(y)|^2 \varrho(y))}^2 .
\end{aligned}$$

For the Sobolev norm, we have to sum over α and interchange the sums, which yields

$$\begin{aligned}\|f \circ \mathrm{R}^{-1}\|_{H^m(\mathbb{T})}^2 &= \|f\|_{L_2(\Omega,\varrho)}^2 + \sum_{\alpha=1}^{m} \left\| \frac{\mathrm{d}^\alpha}{\mathrm{d}x^\alpha} f \circ \mathrm{R}^{-1}(x) \right\|_{L_2(\mathbb{T})}^2 \\ &\leqslant \|f\|_{L_2(\Omega,\varrho)}^2 + \sum_{\alpha=1}^{m} \sum_{k=1}^{\alpha} \left\| \mathrm{D}^k f(y) \right\|_{L_2(\Omega,|B_{\alpha,k}(y)|^2 \varrho(y))}^2 \\ &= \|f\|_{L_2(\Omega,\varrho)}^2 + \sum_{k=1}^{m} \sum_{\alpha=k}^{m} \left\| \mathrm{D}^k f(y) \right\|_{L_2(\Omega,|B_{\alpha,k}(y)|^2 \varrho(y))}^2 .\end{aligned}$$

This motivates us to generalize the Sobolev norm to functions defined on Ω by the following definition.

Definition 4.1 For $m \in \mathbb{N}$, we define the function space

$$H^m(\Omega,\varrho) := \left\{ f : \Omega \to \mathbb{C} \mid \|f\|_{H^m(\Omega,\varrho)} < \infty \right\},$$

where the norm is defined by

$$\|f\|_{H^m(\Omega,\varrho)}^2 := \sum_{k=0}^{m} \left\| \mathrm{D}^k f \right\|_{L_2(\Omega,\Upsilon_{m,k})}^2 \tag{4.3}$$

and the density $\Upsilon_{m,k}$ is defined by

$$\Upsilon_{m,k}(y) := \begin{cases} \sum_{\alpha=k}^{m} |B_{\alpha,k}(y)|^2 \varrho(y) & \text{if } 1 \leqslant k \leqslant m, \\ \varrho(y) & \text{if } k = 0. \end{cases} \tag{4.4}$$

Note that we have for $m \geqslant 2$ the useful recursion formula,

$$\Upsilon_{m,k}(y) = \Upsilon_{m-1,k}(y) + |B_{m,k}(y)|^2 \varrho(y).$$

We state the previous definition for the cases $1 \leqslant m \leqslant 3$ explicitly:

$$\begin{aligned}\|f\|_{H^1(\Omega,\varrho)}^2 &= \|f\|_{L_2(\Omega,\varrho)}^2 + \|f'\|_{L_2(\Omega,1/\varrho)}^2 \\ \|f\|_{H^2(\Omega,\varrho)}^2 &= \|f\|_{L_2(\Omega,\varrho)}^2 + \|f'\|_{L_2(\Omega,1/\varrho+(\varrho')^2/\varrho^5)}^2 + \|f''\|_{L_2(\Omega,1/\varrho^3)}^2 \\ \|f\|_{H^3(\Omega,\varrho)}^2 &= \|f\|_{L_2(\Omega,\varrho)}^2 + \|f'\|_{L_2(\Omega,1/\varrho+(\varrho')^2/\varrho^5+(\varrho'')^2/\varrho^7-6\varrho''(\varrho')^2/\varrho^8+9(\varrho')^4/\varrho^9)}^2 \\ &\quad + \|f''\|_{L_2(\Omega,1/\varrho^3+9(\varrho')^2/\varrho^7)}^2 + \|f'''\|_{L_2(\Omega,1/\varrho^5)}^2 .\end{aligned}$$

This can also be interpreted as follows: the function f can not have a large L_2-norm of its derivatives up to order m in areas where ϱ is small. One can not expect to capture such functions, since where the density is low, we can not approximate derivatives of the function f well. Note that only in special cases, where the derivatives of the density ϱ and the density itself are bounded from below and above, this defined norm is equivalent to a norm defined using the derivatives of a function, weighted with the density ϱ.

Lemma 4.2 *Let $m \in \mathbb{N}$ be positive and the density ϱ fulfill $0 < c_1 \leqslant \|\varrho^{(i)}\|_{L_\infty(\Omega)} \leqslant c_2 < \infty$ for $i = 0, \ldots, m-1$. Then, the norm in (4.3) is equivalent to the norm*

$$\|f\|_{H^m(\Omega,\varrho)} = \sum_{k=0}^{m} \left\| \mathrm{D}^k f \right\|_{L_2(\Omega,\varrho)}.$$

Proof Since every derivative $\mathrm{D}^k \mathrm{R}^{-1}$ can be expressed in terms of derivatives of ϱ up to order $k-1$ and we also assume that ϱ itself is bounded from above and below, the terms $\Upsilon_{m,k}(y)$ can be bounded by

$$0 < C\varrho(y) \leqslant \Upsilon_{m,k}(y) \leqslant C'\varrho(y) < \infty,$$

with some constants $0 < C, C' < \infty$. This yields the assertion. □

Example 4.3 We give three examples for distributions on $\Omega = \mathbb{R}$. We plotted the density function, the transformation R, and the densities $\Upsilon_{m,k}(y)$ from (4.4) in Figs. 3, 4, and 5.

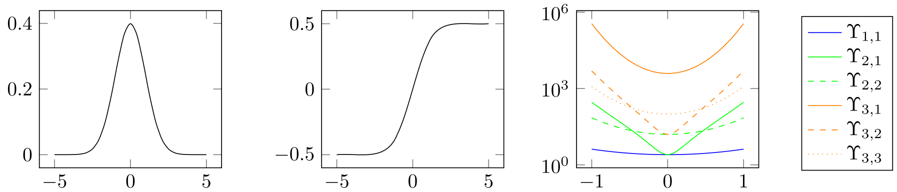
i) **Standard normal distribution on $\mathbb{R}$**
The density

$$\varrho_N(y) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-y^2/2} \tag{4.5}$$

is the *standard normal distribution*. Because of this very smooth density, we expect that this transformation passes the smoothness of f to $f \circ \mathrm{R}^{-1}$. The corresponding transformation is

$$\mathrm{R}(y) = \frac{1}{2}\,\mathrm{erf}\left(\frac{y}{\sqrt{2}}\right), \quad \text{where} \quad \mathrm{erf}(y) = \frac{2}{\sqrt{\pi}} \int_y^\infty \mathrm{e}^{-t^2}\,\mathrm{d}t.$$

ii) **Cauchy distribution on $\mathbb{R}$**

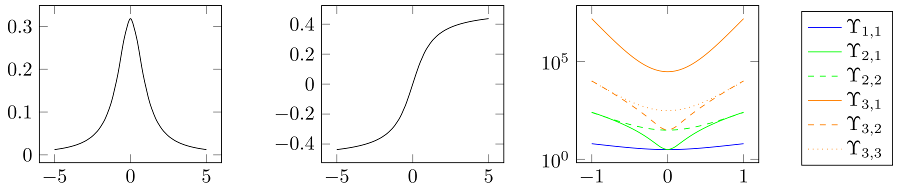


(a) Density function $\varrho_N(y)$.

(b) The transformation $\mathrm{R}(y)$.

(c) The functions $\Upsilon_{m,k}(y)$ defined in (4.4).

Fig. 3 The standard normal distribution ϱ_N on $\mathbb{R}$



(a) Density function $\varrho_C(y)$.

(b) The transformation $\mathrm{R}(y)$.

(c) The functions $\Upsilon_{m,k}(y)$ defined in (4.4).

Fig. 4 The Cauchy distribution ϱ_C on $\mathbb{R}$

The density

$$\varrho_C(y) = \frac{1}{\pi\,(1+y^2)} \tag{4.6}$$

is a *Cauchy distribution*. The corresponding transformation is

$$\mathrm{R}(y) = \frac{1}{\pi} \arctan y.$$

iii) **Laplace distribution** $\mathbb{R}$

The density

$$\varrho_L(y) = \frac{1}{8} \mathrm{e}^{-\frac{|y-2|}{4}} \tag{4.7}$$

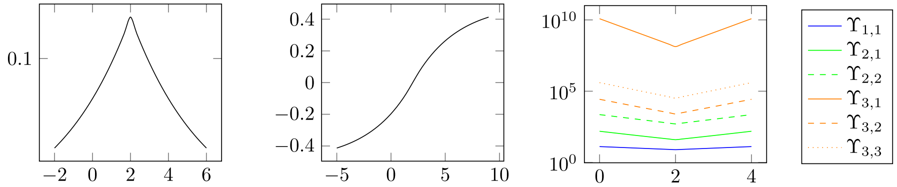
is a *Laplace distribution*, and the corresponding transformation is

$$\mathrm{R}(y) = \frac{1}{2}\,\mathrm{sgn}(y-2)\left(1 - \mathrm{e}^{-\frac{|y-2|}{4}}\right).$$

So far, we characterized for natural numbers m function spaces where the transformation R preserves the smoothness. The definition of the Sobolev norm using the decay of the Fourier coefficients in (A.2) allows us to study functions of fractional smoothness. Hence, we define fractional smoothness for functions defined on Ω.

Definition 4.4 Let $s > 0$. Then, we define

$$H^s(\Omega, \varrho) := \left\{ f : \Omega \to \mathbb{C} \mid \|f\|_{H^s(\Omega,\varrho)} < \infty \right\},$$



(a) Density function $\varrho_L(y)$.

(b) The transformation $\mathrm{R}(y)$.

(c) The functions $\Upsilon_{m,k}(y)$ defined in (4.4).

Fig. 5 The Laplace distribution ϱ_L on $\mathbb{R}$

where the norm is defined by

$$\|f\|_{H^s(\Omega,\varrho)}^2 := \sum_{k\in\mathbb{Z}} |c_k^{\varrho}(f)|^2(1+|k|^2)^s$$

with the Fourier coefficients $c_k^{\varrho}(f)$ of the transformed function $c_k^{\varrho}(f) := c_k(f \circ \mathrm{R}^{-1})$ and the Fourier coefficients for periodic functions are defined in (2.1).

Remark 4.5 The norm in the previous definition is for $m = s \in \mathbb{N}_0$ equivalent to the norm from Definition 4.1, since the terms $\Upsilon_{m,k}(y)$ are chosen such that

$$\sum_{k=1}^{m} \|\mathrm{D}^k(f \circ \mathrm{R}^{-1})\|_{L_2(\mathbb{T})}^2 = \sum_{k=1}^{m} \left\|\mathrm{D}^k f\right\|_{L_2(\Omega,\Upsilon_{m,k})}^2,$$

because of the norm equality of the norms (A.1) and (A.2), see [24].

The multivariate setting

The theory from the one-dimensional case can be transferred to the d-dimensional setting. Again, we have to use (A.1) and have to estimate norms of derivatives of the transformed function $f \circ \mathrm{R}^{-1}$. Using the equations (3.9), we have that

$$\mathrm{D}^{\boldsymbol{\alpha}}(f \circ \mathrm{R}^{-1})(\boldsymbol{x}) = \sum_{\boldsymbol{k}=\boldsymbol{1}}^{\boldsymbol{\alpha}} B_{\boldsymbol{\alpha},\boldsymbol{k}}(\boldsymbol{y})\, \mathrm{D}^{\boldsymbol{k}} f(\boldsymbol{y}), \tag{4.8}$$

where we define the multivariate analogon to (4.1) by

$$B_{\boldsymbol{\alpha},\boldsymbol{k}}(\boldsymbol{y}) := \prod_{i=1}^{d} B_{\alpha_i,k_i}(y_{k_i}). \tag{4.9}$$

This motivates to generalize Definition 4.1 to multivariate Sobolev spaces of mixed dominating smoothness by

Definition 4.6 For $m \in \mathbb{N}$ and $\Omega_i \in \{\mathbb{T}, \mathbb{R}\}$, we define the function space

$$H_{\text{mix}}^m(\Omega, \varrho) := \left\{ f : \Omega \to \mathbb{C} \mid \|f\|_{H_{\text{mix}}^m(\Omega,\varrho)} < \infty \right\},$$

where the norm is defined by

$$\|f\|_{H_{\text{mix}}^m(\Omega,\varrho)}^2 = \sum_{0 \leqslant \|\boldsymbol{k}\|_\infty \leqslant m} \left\|\mathrm{D}^{\boldsymbol{k}} f(\boldsymbol{y})\right\|_{L_2(\Omega,\Upsilon_{m,\boldsymbol{k}})}^2$$

and the density $\Upsilon_{m,\boldsymbol{k}}(y)$ is defined by

$$\Upsilon_{m,\boldsymbol{k}}(\boldsymbol{y}) := \prod_{i=1}^{d} \Upsilon_{m,k_i}(y_{k_i}),$$

where the one-dimensional functions Υ_{m,k_i} are defined in (4.4).

This $H^m_{\text{mix}}(\Omega, \varrho)$-norm is equivalent to a norm definition using the decay of the Fourier coefficients of the transformed function $c^{\varrho}_{\boldsymbol{k}}(f \circ \mathrm{R}^{-1})$ like in (A.2), which can be generalized to the case of fractional smoothness by

$$H^s_{\text{mix}}(\Omega, \varrho) := \left\{ f : \Omega \to \mathbb{C} \mid \|f\|_{H^s_{\text{mix}}(\Omega,\varrho)} < \infty \right\}, \tag{4.10}$$

where the norm is defined by

$$\|f\|^2_{H^s_{\text{mix}}(\Omega,\varrho)} = \sum_{\boldsymbol{k}\in\mathbb{Z}^d} |c^{\varrho}_{\boldsymbol{k}}(f)|^2 \prod_{i=1}^{d} (1 + |k_i|^2)^s. \tag{4.11}$$

The function spaces $H^s_{\text{mix}}(\Omega, \varrho)$ are defined such that the transformed function $f \circ \mathrm{R}^{-1}$ inherits the smoothness of the function f, i.e.,

$$\|f\|_{H^s_{\text{mix}}(\Omega,\varrho)} \lesssim \|f \circ \mathrm{R}^{-1}\|_{H^s_{\text{mix}}(\mathbb{T}^d)} . \tag{4.12}$$

4.2 Weighted function spaces in the literature

There is a huge literature about weighted function spaces. We restrict ourselves to a few references closely related to our approach. Sobolev norms as defined in Definition 4.1, where the norms of the derivatives are measured with respect to a different density are also considered in [14] in the case $m = 1$. For the one-dimensional case on $\Omega = \mathbb{R}$, the authors showed that the norms $\|f\|^2_{\mathcal{H}} = \|f\|^2_{L_2(\mathbb{R},\varrho)} + \frac{1}{\gamma}\,\|\mathrm{D}^1 f\|^2_{L_2(\mathbb{R},\psi)}$ and $\|f\|^2_{\mathcal{W}} = |\int_{-\infty}^{\infty} f(y)\varrho(y)\,\mathrm{d}y|^2 + \frac{1}{\gamma}\,\|\mathrm{D}^1 f\|^2_{L_2(\mathbb{R},\psi)}$ are equivalent under certain conditions on the density ψ. We are in the special case where $\psi(y) = {}^1\!/_{\varrho(y)}$ and meet the conditions in our examples. The authors also showed a norm equivalence for multivariate functions.

Another example of weighted norms is [30], where the authors weight the derivatives of functions with some exponential term in order to integrate functions on $\mathbb{R}$ numerically.

In [40], several weighted function spaces were introduced. On $\mathbb{R}^d$, the authors defined the weighted counterpart to the classical Sobolev spaces by introducing a weight function w for weighting all derivatives of the function f similarly, i.e., in our notation

$$\|f\|_{W^m_2(\mathbb{R}^d,w)} := \left(\sum_{|\boldsymbol{\alpha}|_1 \leqslant m} \|\mathrm{D}^{\boldsymbol{\alpha}} f\|^2_{L_2(\mathbb{R}^d,w^2)} \right)^{1/2} .$$

An admissible weight function w has the following properties (see [40, Definition 6.1]).

Definition 4.7 The class W^d of admissible weight functions w is the collection of all positive functions $w \in C^{\infty}(\mathbb{R}^d)$ with the following properties:

i) For all $\boldsymbol{\gamma} \in \mathbb{N}^d_0$, there is a positive constant $c_{\boldsymbol{\gamma}}$ with

$$|\mathrm{D}^{\boldsymbol{\gamma}} w(\boldsymbol{y})| \leqslant c_{\boldsymbol{\gamma}} w(\boldsymbol{y}) \text{ for all } \boldsymbol{y} \in \mathbb{R}^d. \tag{4.13}$$

i) There are two constants $c > 0$ and $\alpha \geqslant 0$ such that

$$0 < w(\boldsymbol{y}_1) \leqslant c w(\boldsymbol{y}_2) \left(1 + |\boldsymbol{y}_1 - \boldsymbol{y}_2|^2\right)^{\alpha/2} \quad \text{for all } \boldsymbol{y}_1 \in \mathbb{R}^d, \boldsymbol{y}_2 \in \mathbb{R}^d.$$

We have the following connection between these weighted spaces $W_2^m(\mathbb{R}^d, w)$ and the spaces $H^m(\mathbb{R}^d, \varrho)$ defined in this paper.

Theorem 4.8 *Let $m \in \mathbb{N}$, $\varrho \in W^d$ be an admissible weight function with $\varrho(\boldsymbol{y}) < \infty$ for all $\boldsymbol{y} \in \mathbb{R}^d$. Then*

$$W_2^m(\mathbb{R}^d, \varrho^{-m+1/2}) \subset H^m(\mathbb{R}^d, \varrho).$$

Proof We begin with the one-dimensional case. First, we show that for the Bell polynomials defined in (4.1), there holds

$$|B_{\alpha,k}(y)| \lesssim |\varrho(y)|^{-\alpha}$$

by induction over α. The condition (4.13) of an admissible weight function and the examples (4.2) show that this is true for $\alpha \leqslant 3$. The Bell polynomials can be described recursively by [4]

$$B_{\alpha,k}(y) = \sum_{j=1}^{\alpha-k+1} \binom{\alpha-1}{j-1} B_{\alpha-1,k-1}((\mathrm{R}^{-1})^{(1)}(x), (\mathrm{R}^{-1})^{(2)}(x), \ldots, (\mathrm{R}^{-1})^{(\alpha-j-k+2)}(x))(\mathrm{R}^{-1})^{(j)}(\mathrm{R}(y)). \tag{4.14}$$

The derivatives of R^{-1} can be bounded by

$$|(\mathrm{R}^{-1})^{(j)}(\mathrm{R}(y))| \lesssim \frac{1}{\varrho(y)}, \quad \text{for } j \geqslant 1 \tag{4.15}$$

since $(\mathrm{R}^{-1})^{(1)}(\mathrm{R}(y)) = \frac{1}{\varrho(y)}$ and inductively every further derivative (of the sum of several terms which are fractions of polynomials of derivatives of ϱ and a power of ϱ in the denominator) either increases the power of ϱ in the denominator while adding a ϱ' in the nominator or just increases the derivatives of ϱ (but not the degree of the polynomial) in the nominator. The condition (4.13) shows than (4.15), which gives by induction and using (4.14) the result (4.15). This again gives us that

$$|\Upsilon_{m,k}(y)|^2 \lesssim \frac{1}{(\varrho(y))^{2m-1}},$$

which is also true in case where $m = 0$ because of the condition that $\varrho(y) < \infty$. The choice $w = \varrho^{-m+1/2} \in W^d$ gives then

$$\|\mathrm{D}^k f\|_{L_2(\mathbb{R}^d, w^2)} \leqslant \|\mathrm{D}^k f\|_{L_2(\mathbb{R}^d, \Upsilon_{m,k})}$$

for $0 \leqslant k \leqslant m$. The multivariate case follows from the fact that ϱ is assumed to be a product density, such that the weight w is also a product weight, since the densities $\Upsilon_{m,\boldsymbol{k}}(\boldsymbol{y})$, defined in (4.11), are also products of the one-dimensional functions. □

Note that, for instance, the constant function is in all $H^m(\mathbb{R}^d, \varrho)$, but not in $W^m(\mathbb{R}^d, \varrho^{-m+1/2})$, since $\lim_{x\to\pm\infty} \varrho(x) = 0$. The Cauchy distribution (4.6) belongs to the set of admissible weight functions. It is also possible to extend this theory of weighted function spaces to other weight functions, for instance, exponential weights. Then, one has to change the definition of admissible weights (see [40, Remark 6.4]), but the connection to our function spaces via Theorem 4.8 is nevertheless possible.

4.3 Weighted Besov spaces

So far, we studied the smoothness of the transformed function $f \circ \mathrm{R}$ in Sobolev spaces with dominating mixed derivatives. The advantage of Besov-Nikolskij spaces compared to the Sobolev spaces is that they are a much more general tool in describing the smoothness properties of functions.

Be aware that we will use in this chapter the indices $\boldsymbol{j}$ and $\boldsymbol{k}$ not as wavelet indices, but $\boldsymbol{k}$ as index of the Fourier coefficients and $\boldsymbol{j}$ as index of dyadic blocks in which we decompose the indices $\boldsymbol{k}$.

We use the Fourier analytic characterization of the spaces (A.3). Therefore, we introduce the dyadic blocks

$$J_j = \begin{cases} \{k \in \mathbb{Z} \mid 2^{j-1} \leqslant |k| < 2^j\} & \text{if } j \geqslant 1, \\ \{0\} & \text{if } j = 0. \end{cases}$$

For $\boldsymbol{j} \in \mathbb{N}_0^d$, we define

$$J_{\boldsymbol{j}} := J_{j_1} \times \ldots \times J_{j_d},$$

if all components belong to $\mathbb{N}_0$. Using these dyadic blocks, we decompose the Fourier series of the function f into

$$f = \sum_{\boldsymbol{j}\in\mathbb{N}_0^d} f_{\boldsymbol{j}}(\boldsymbol{x}) \text{ with } f_{\boldsymbol{j}}(\boldsymbol{x}) := \sum_{\boldsymbol{k}\in J_{\boldsymbol{j}}} c_{\boldsymbol{k}}(f)\mathrm{e}^{2\pi\mathrm{i}\langle\boldsymbol{k},\boldsymbol{x}\rangle}. \tag{4.16}$$

Furthermore, we introduce the Fourier coefficients for functions defined on Ω, using the Fourier coefficients of the transformed function $f \circ \mathrm{R}^{-1}$ from (2.1) by

$$c_{\boldsymbol{k}}^{\varrho}(f) = \int_{\mathbb{T}^d} (f \circ \mathrm{R}^{-1})(\boldsymbol{x})\, \mathrm{e}^{-2\pi\mathrm{i}\langle\boldsymbol{k},\boldsymbol{x}\rangle}\, \mathrm{d}\boldsymbol{x} = \int_{\Omega} f(\boldsymbol{y})\mathrm{e}^{-2\pi\mathrm{i}\langle\boldsymbol{k},\mathrm{R}(\boldsymbol{y})\rangle}\varrho(\boldsymbol{y})\, \mathrm{d}\boldsymbol{y}.$$

Therefore,

$$f(\boldsymbol{y}) = \sum_{\boldsymbol{k}\in\mathbb{Z}^d} c_{\boldsymbol{k}}^{\varrho}(f)\, \mathrm{e}^{2\pi\mathrm{i}\langle\mathrm{R}(\boldsymbol{y}),\boldsymbol{k}\rangle}.$$

This yields immediately by (3.10).

$$\sum_{\boldsymbol{k}\in\mathbb{Z}^d} |c_{\boldsymbol{k}}^{\varrho}(f)|^2 = \|f\|^2_{L_2(\Omega,\varrho)}\,.$$

For the Definition in (A.3), we have to split the periodic function $f \circ \mathrm{R}^{-1}$ in dyadic blocks belonging to the indices $\boldsymbol{j}$. For a fixed index $\boldsymbol{j} \in \mathbb{N}_0^d$, we define $\boldsymbol{u} = \{i \in [d] \mid j_i > 0\}$. Similar to the decomposition with wavelet functions, one can describe a connection of Fourier coefficients and ANOVA terms (see [31]) for a function $g \in L_2(\mathbb{T}^d)$

$$c_{\boldsymbol{k}}(g_{\boldsymbol{u}}) \neq 0 \Leftrightarrow \operatorname{supp} \boldsymbol{k} := \{i \in [d] \mid k_i \neq 0\} = \boldsymbol{u}.$$

and

$$c_{\boldsymbol{k}}(g) = c_{\boldsymbol{k}_{\boldsymbol{u}}}(g_{\operatorname{supp} k}). \tag{4.17}$$

The splitting of the function $f \circ \mathrm{R}^{-1}$ in dyadic blocks $\boldsymbol{j} \neq \boldsymbol{0}$ gives for

$$\boldsymbol{\alpha} \in \{\alpha \in \mathbb{N}_0^d \mid \operatorname{supp} \boldsymbol{\alpha} = \boldsymbol{u}, \sum_{i \in \boldsymbol{u}} \alpha_i = m\},$$

using $|c_k(g)| = \frac{1}{k} c_k(g')$ and (4.8) that

$$\begin{aligned}
\|(f \circ \mathrm{R}^{-1})_{\boldsymbol{j}}\|_{L_2(\mathbb{T}^d)}^2 &= \sum_{\boldsymbol{k} \in J_{\boldsymbol{j}}} |c_{\boldsymbol{k}}(f \circ \mathrm{R}^{-1})|^2 \overset{(4.17)}{=} \sum_{\boldsymbol{k} \in J_{\boldsymbol{j}}} |c_{\boldsymbol{k}_{\boldsymbol{u}}}((f \circ \mathrm{R}^{-1})_{\boldsymbol{u}})|^2 \\
&= \sum_{\boldsymbol{k} \in J_{\boldsymbol{j}}} \frac{1}{\prod_{i \in \boldsymbol{u}} |k_i|^{2\alpha_i}} \left| c_{\boldsymbol{k}_{\boldsymbol{u}}} \left(\frac{\mathrm{d}^{\boldsymbol{\alpha}_{\boldsymbol{u}}}}{\mathrm{d}x^{\boldsymbol{\alpha}_{\boldsymbol{u}}}} (f \circ \mathrm{R}^{-1})_{\boldsymbol{u}} \right) \right|^2 \\
&\overset{(4.8)}{=} \sum_{\boldsymbol{k} \in J_{\boldsymbol{j}}} \frac{1}{\prod_{i \in \boldsymbol{u}} |k_i|^{2\alpha_i}} \left| c_{\boldsymbol{k}_{\boldsymbol{u}}} \left(\sum_{\boldsymbol{\beta}=\boldsymbol{1}}^{\boldsymbol{\alpha}_{\boldsymbol{u}}} B_{\boldsymbol{\alpha}_{\boldsymbol{u}},\boldsymbol{\beta}}(\boldsymbol{y}_{\boldsymbol{u}}) \, \mathrm{D}^{\boldsymbol{\beta}} f_{\boldsymbol{u}}(\boldsymbol{y}_{\boldsymbol{u}}) \right) \right|^2 \\
&\leqslant \sum_{\boldsymbol{\beta}=\boldsymbol{1}}^{\boldsymbol{\alpha}_{\boldsymbol{u}}} \sum_{\boldsymbol{k} \in J_{\boldsymbol{j}}} \frac{1}{\prod_{i \in \boldsymbol{u}} |k_i|^{2\alpha_i}} \left| \int_{\mathbb{T}} B_{\boldsymbol{\alpha}_{\boldsymbol{u}},\boldsymbol{\beta}}(\mathrm{R}_{\boldsymbol{u}}^{-1}(\boldsymbol{x}_{\boldsymbol{u}})) \, \mathrm{D}^{\boldsymbol{\beta}} f(\mathrm{R}_{\boldsymbol{u}}^{-1}(\boldsymbol{x}_{\boldsymbol{u}})) \mathrm{e}^{-2\pi \mathrm{i} \langle \boldsymbol{x}_{\boldsymbol{u}}, \boldsymbol{k}_{\boldsymbol{u}} \rangle} \, \mathrm{d}\boldsymbol{x}_{\boldsymbol{u}} \right|^2 \\
&= \sum_{\boldsymbol{\beta}=\boldsymbol{1}}^{\boldsymbol{\alpha}_{\boldsymbol{u}}} \sum_{\boldsymbol{k} \in J_{\boldsymbol{j}}} \frac{1}{\prod_{i \in \boldsymbol{u}} |k_i|^{2\alpha_i}} \left| \int_{\Omega} B_{\boldsymbol{\alpha}_{\boldsymbol{u}},\boldsymbol{\beta}}(\boldsymbol{y}_{\boldsymbol{u}}) \, \mathrm{D}^{\boldsymbol{\beta}} f_{\boldsymbol{u}}(\boldsymbol{y}_{\boldsymbol{u}}) \mathrm{e}^{-2\pi \mathrm{i} \langle \mathrm{R}_{\boldsymbol{u}}(\boldsymbol{y}_{\boldsymbol{u}}), \boldsymbol{k} \rangle} \varrho_{\boldsymbol{u}}(\boldsymbol{y}_{\boldsymbol{u}}) \, \mathrm{d}\boldsymbol{y}_{\boldsymbol{u}} \right|^2 \\
&= \sum_{\boldsymbol{\beta}=\boldsymbol{1}}^{\boldsymbol{\alpha}_{\boldsymbol{u}}} \sum_{\boldsymbol{k} \in J_{\boldsymbol{j}}} \frac{1}{\prod_{i \in \boldsymbol{u}} |k_i|^{2\alpha_i}} \left| c_{\boldsymbol{k}_{\boldsymbol{u}}}^{\varrho_{\boldsymbol{u}}} \left(B_{\boldsymbol{\alpha}_{\boldsymbol{u}},\boldsymbol{\beta}}(\boldsymbol{y}_{\boldsymbol{u}}) \, \mathrm{D}^{\boldsymbol{\beta}} f_{\boldsymbol{u}}(\boldsymbol{y}_{\boldsymbol{u}}) \right) \right|^2,
\end{aligned}$$

where we used the notation from (4.9) and $\beta \in \mathbb{N}^{|\boldsymbol{u}|}$. In the special case $\boldsymbol{j} = \boldsymbol{0}$, we have $\boldsymbol{k} = \boldsymbol{0}$ and

$$\|(f \circ \mathrm{R}^{-1})_{\boldsymbol{j}}\|_{L_2(\mathbb{T}^d)}^2 = \int_{\mathbb{T}^d} f \circ \mathrm{R}^{-1}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = f_{\varnothing}$$

Therefore, we define a Besov norm for functions defined on Ω by

Definition 4.9 Let $s > 1/2$ and $m = \lfloor s \rfloor$. Then, we define

$$\|f\|_{\boldsymbol{B}_{2,\infty}^s(\Omega,\varrho)}^2 := \max \left\{ f_{\varnothing}, \sup_{\substack{\boldsymbol{j} \in \mathbb{N}^d \\ \boldsymbol{j} \neq \boldsymbol{0}}} 2^{2|\boldsymbol{j}|_1 s} \sup_{\substack{\boldsymbol{\alpha}_{\boldsymbol{u}} \in \mathbb{N}^{|\boldsymbol{u}|} \\ |\boldsymbol{\alpha}_{\boldsymbol{u}}| = m}} \sup_{\boldsymbol{1} \leqslant \boldsymbol{\beta} \leqslant \boldsymbol{\alpha}_{\boldsymbol{u}}} \sum_{\boldsymbol{k} \in J_{\boldsymbol{j}}} \frac{1}{\prod_{i \in \boldsymbol{u}} |k_i|^{2\alpha_i}} \left| c_{\boldsymbol{k}_{\boldsymbol{u}}}^{\varrho_{\boldsymbol{u}}} \left(B_{\boldsymbol{\alpha}_{\boldsymbol{u}},\boldsymbol{\beta}}(\boldsymbol{y}_{\boldsymbol{u}}) \, \mathrm{D}^{\boldsymbol{\beta}} f_{\boldsymbol{u}}(\boldsymbol{y}_{\boldsymbol{u}}) \right) \right|^2 \right\}.$$

This definition yields the estimate

$$\|f \circ \mathrm{R}^{-1}\|_{\boldsymbol{B}^s_{2,\infty}(\mathbb{T}^d)} \lesssim \|f\|_{\boldsymbol{B}^s_{2,\infty}(\Omega,\varrho)} \,. \tag{4.18}$$

In contrast to the norm $\|f \circ \mathrm{R}^{-1}\|_{\boldsymbol{B}^s_{2,\infty}(\mathbb{T}^d)}$, which deals with the transformed function on $\mathbb{T}^d$, the norm defined in Definition 4.9 considers products of the function f with terms consisting of powers and derivatives of the density ϱ.

4.4 A note on non-periodic functions

So far, we studied function spaces for the case $\Omega_i \in \{\mathbb{T}, \mathbb{R}\}$ for all $i \in [d]$. A key difficulty in the approximation on $\Omega = [0, 1]$ is the non-periodicity of the function, i.e., the behavior at the boundary. No matter how the density ϱ looks like, a transformation which equals the cumulative distribution function can not ensure that the function $f \circ \mathrm{R}^{-1}$ is periodic. Therefore, we introduced in the transformation (3.2) the extension parameter η. We denote the extension of the function f by $\tilde{f}$ with $\tilde{f}\big|_{[-1/2+\eta, 1/2]} = f \circ \mathrm{R}^{-1}$. The L_2-norm of the function f itself behaves like the L_2-norm of the transformation up to some factor (see (3.10)). The same is true for the derivatives. For $\alpha \in \mathbb{N}$ with $\alpha \leqslant m$, we have

$$\begin{aligned}
\|\tilde{f}^{(\alpha)}\|^2_{L_2([0,1])} &= \|\mathrm{D}^{(\alpha)}\tilde{f}\|^2_{L_2([-1/2,-1/2+\eta])} + \frac{1}{(1-\eta)}\sum_{k=1}^{\alpha}\|\mathrm{D}^k f\|^2_{L_2([0,1],|B_{\alpha,k}(y)|^2\varrho(y))} \\
&= \|\mathrm{D}^{(\alpha)}\tilde{f}\|^2_{L_2([-1/2,-1/2+\eta])} + \sum_{k=1}^{\alpha}\|\mathrm{D}^k f\|^2_{L_2([0,1],|B_{\alpha,k}(y)|^2\frac{\varrho(y)}{(1-\eta)})} \,.
\end{aligned} \tag{4.19}$$

Hence, for every non-periodic direction, we get the additional factor $(1-\eta)$. This gives us the reasonable assumption that extension $\tilde{f}$ at the boundary has to have a small Sobolev norm. For more details, see Sect. 5.1. Since the factor $(1-\eta)$ is a constant for fixed η, we define for the sake of simplicity for function on $[0, 1]^d$ the weighted function spaces in the same manner as for periodic functions with the following definition.

Definition 4.10 For $m \in \mathbb{N}$, let the extension of $f \circ \mathrm{R}^{-1}$ to the boundary be in the Sobolev space $H^m_{\mathrm{mix}}(\mathbb{T}^d \backslash [-1/2+\eta, 1/2]^d)$. Then, we define the function space

$$H^m_{\mathrm{mix}}([0,1]^d,\varrho) := \left\{ f : [0,1]^d \to \mathbb{C} \mid \|f\|_{H^m_{\mathrm{mix}}([0,1]^d,\varrho)} < \infty \right\},$$

where the norm is defined by

$$\|f\|^2_{H^m_{\mathrm{mix}}([0,1]^d,\varrho)} = \sum_{0 \leqslant \|\boldsymbol{k}\|_\infty \leqslant m} \left\| \mathrm{D}^{\boldsymbol{k}} f(\boldsymbol{y}) \right\|^2_{L_2([0,1]^d,\Upsilon_{m,\boldsymbol{k}})}$$

and the density $\Upsilon_{m,\boldsymbol{k}}(y)$ is defined by

$$\Upsilon_{m,\boldsymbol{k}}(\boldsymbol{y}) := \prod_{i=1}^{d} \Upsilon_{m,k_i}(y_{k_i}),$$

where the one-dimensional functions Υ_{m,k_i} are defined in (4.4).

Of course, this definition can be mixed with Definition 4.6 for a mixed function, which has different domains Ω_i for different $i \in [d]$. Analogously, the function spaces for fractional smoothness s and for mixed Besov spaces are defined like in the periodic case.

If the function f has certain smoothness m on the interval $[0, 1]$, the transformed function inherits the smoothness of this function, as in the cases for periodic functions. This means, if the Sobolev norm of the extended function $\tilde{f}$ at boundary is finite, we have

$$f \in H^s_{\mathrm{mix}}([0,1]^d, \varrho) \Leftrightarrow f \circ \mathrm{R}^{-1} \in H^s_{\mathrm{mix}}([-\tfrac{1}{2} + \eta, \tfrac{1}{2}]^d) \Leftrightarrow \tilde{f} \in H^s_{\mathrm{mix}}(\mathbb{T}^d).$$

Let us finish this excursion to the non-periodic setting with an example for the distribution ϱ.

Example 4.11 Beta distribution on the interval $[0, 1]$

Let $\Omega = [0, 1]$ be the unit interval. For $\alpha > 0$, we define by

$$\varrho_{B,\alpha}(y) = \frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2} y^{\alpha-1}(1-y)^{\alpha-1} \tag{4.20}$$

the *beta distribution* with the shape parameter α, where Γ be the Gamma function. For $\alpha = 1$, this is the uniform distribution. For $\alpha > 1$, the cumulative distribution function is the regularized incomplete beta function, so the transformation in the case $\eta = 0$ reads

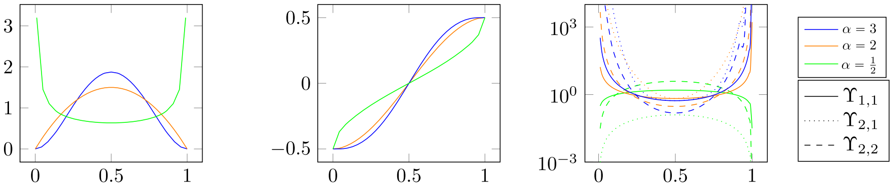
$$\mathrm{R}(y) = \frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2} \int_0^y (t^2 - t)^{\alpha-1}\,\mathrm{d}t - \tfrac{1}{2},$$

which can be computed analytically for fixed α. These functions are plotted in Fig. 6 for different parameters α.

Remark 4.12 The beta distribution $\varrho_{B,1/2}$ coincides with the Chebychev distribution, which is defined on $[-1, 1]$. In this case [22, Section 10.3], [13] propose the Chebyshev polynomials as basis in $L_2([-1,1]^d, \varrho_{B,1/2})$. This coincides with using our transformation R and the cosine basis on $\mathbb{T}^d$.

5 First setting: known density ρ

In this chapter, we study the case where we assume that the underlying density ϱ of the samples is known, and it is a tensor product density (3.7). With this information, we use the transformation (3.2), transform the given samples $\mathcal{Y} \subset \Omega$ to the transformed



(a) Density function $\varrho_{B,\alpha}(y)$. (b) The transformation $\mathrm{R}(y)$ for $\eta = 0$. (c) The functions $\Upsilon_{m,k}(y)$ defined in (4.4).

Fig. 6 The beta distribution $\varrho_{B,\alpha}$ on the interval $[0, 1]$ for $\alpha \in \{1/2, 2, 3\}$

samples $\mathcal{X} = \mathrm{R}(\mathcal{Y})$ on the torus, and apply the approximation operator $S_n^{\mathcal{X}}$, given in (2.3). With the introduction of weighted function spaces in the previous chapter, we estimate the error of this approximation if the function f itself is in a weighted function space. In fact, we formulate the following theorem.

Theorem 5.1 *Let $\Omega_i \in \{\mathbb{T}, \mathbb{R}\}$ the density ϱ_i be in $C^{m-1}(\Omega_i)$ for $i \in [d]$, and let $m \in \mathbb{N}$ be the order of vanishing moments of the wavelet ψ. Let the function fulfill for all $i \in [d]$ where $\Omega_i = \mathbb{R}$ that*

$$\lim_{y_i \to \infty} f(\boldsymbol{y}) = \lim_{y_i \to -\infty} f(\boldsymbol{y}).$$

Let furthermore M be the number of samples satisfying $M \gtrsim rN \log N$, where $N = |I_n|$ is the number of wavelet functions and $r > 1$. Let $\mathcal{Y} = (\boldsymbol{y}_j)_{j=1}^M \subset \Omega$ be drawn i.i.d,. at random according to ϱ, $f \in C(\Omega)$ a continuous function and the samples $\mathcal{Y}$ transformed to $\mathcal{X} = \mathrm{R}(\mathcal{Y})$ using (3.7). In the case where $1/2 < s < m$, we have

$$\mathbb{P}\left(\|f - (S_n^{\mathcal{X}}(f \circ \mathrm{R}^{-1})) \circ \mathrm{R}\|_{L_2(\Omega,\varrho)} \lesssim 2^{-ns} n^{(d-1)/2} \, \|f\|_{\boldsymbol{B}^s_{2,\infty}(\Omega,\varrho)}\right) \geqslant 1 - 2\,M^{-r}. \tag{5.1}$$

And in the case $s = m$, we have

$$\mathbb{P}\left(\|f - (S_n^{\mathcal{X}}(f \circ \mathrm{R}^{-1})) \circ \mathrm{R}\|_{L_2(\Omega,\varrho)} \lesssim 2^{-ns} n^{(d-1)/2} \, \|f\|_{H^s_{\mathrm{mix}}(\Omega,\varrho)}\right) \geqslant 1 - 2\,M^{-r}. \tag{5.2}$$

Proof The theory from in [26] studies the behavior of periodic functions on $\mathbb{T}^d$. Because of the assertions, the function $f \circ \mathrm{R}^{-1}$ is a function on $\mathbb{T}^d$, and the samples $\mathcal{X} = \mathrm{R}(\mathcal{Y})$ are uniform i.i.d. Hence, [26, Corollary 3.22] is applicable to the function $f \circ \mathrm{R}^{-1}$. Together with the definitions of the Sobolev and Besov norms for functions of mixed smoothness, for which we have (4.12) as well as (4.18), this yields the assertion. □

This theorem is about the case where we do not have non-periodic variables y_i involved. We introduced in (3.2) the extension parameter η to also deal with non-periodic functions. In the next chapter, we will give a similar result for the non-periodic case. Of course, these results can be mixed if the domain has different parts Ω_i.

5.1 Extensions of non-periodic functions

Here, we study the case where $\Omega = [0, 1]^d$. Of course, one can interpret a function on $[0, 1]^d$ as a (possibly non-continuous) function on the torus by gluing the endpoints together. This coincides with the transformation (3.2) with no extension $\eta = 0$. Since the function $f \circ \mathrm{R}^{-1}$ is then non-continuous, Theorem 5.1 does not give us a reasonable error decay. For that reason, the aim of this section is to show that the transformation idea also works for non-periodic functions with a reasonable choice of the extension parameter η in the transformation (3.2).

For Fourier approximation, there is the approach of *Fourier extension* [3, 5, 20], where the function is continued outside of the interval [0, 1] to a smooth function. See also [2] for a nice overview and the connection to the frame approach. We use a similar approach by introducing the extension parameter η, which allows us to extend $\tilde{f}$ on the boundary $[-1/2, -1/2 + \eta]$ in an appropriate way. On one hand, this gives better approximation rates, but on the other hand, the stability gets worse. The occurring problem is that we have to bound the condition of the approximation matrix $\boldsymbol{A}$ (see (2.2)). To circumvent this, in the mentioned literature, the authors use for instance the truncated singular value decomposition. We do not want to set up the whole matrix $\boldsymbol{A}$, but rather use a least squares algorithm which gets only the result of a matrix-vector multiplication with $\boldsymbol{A}$ and $\boldsymbol{A}^\top$. We will see in this chapter that an appropriate choice of the extension parameter ensures stability.

Remember, rather than taking the cumulative distribution function of the density ϱ, we use the modification

$$\mathrm{R}(y) = \eta + (1-\eta) \int_0^y \varrho(t)\, \mathrm{d}t, \tag{5.3}$$

where $0 < \eta \ll 1$ is some extension parameter. In fact, we get the transformed samples $\mathcal{X} = \mathrm{R}(\mathcal{Y})$, which are uniformly distributed on the cube $\tilde{\Omega} := [-1/2 + \eta, 1/2]^d$. This procedure transforms and compresses the original function f into the box $\tilde{\Omega}$ and allows to extend this function to a function $\tilde{f}$ defined on $\mathbb{T}^d$. Figure 7 shows an illustration of the two-dimensional domains. On the boundary $[-1/2, -1/2 + \eta]$, we extend the function f. As mentioned in the discussion before Definition 4.10, it is reasonable to choose an extension which has Sobolev smoothness. For the following result, we introduce the notation of restricting function spaces $V \subset L_2(\mathbb{T}^d)$ to some smaller domain by

$$V\big|_{[a,b]^d} := \left\{ g\big|_{[a,b]^d} : g \in V \right\}$$

for some cube $[a, b]^d \subset \mathbb{T}^d$.

Lemma 5.2 *Let m be the order of the Chui-Wang wavelets and n the maximal wavelet level. Denote the hyperbolic function spaces by*

$$V_n^{(d)} := \operatorname{span}\{\psi_{\boldsymbol{j},\boldsymbol{k}}^{\mathrm{per}} \mid (\boldsymbol{j}, \boldsymbol{k}) \in I_n\} \subset L_2(\mathbb{T}^d).$$

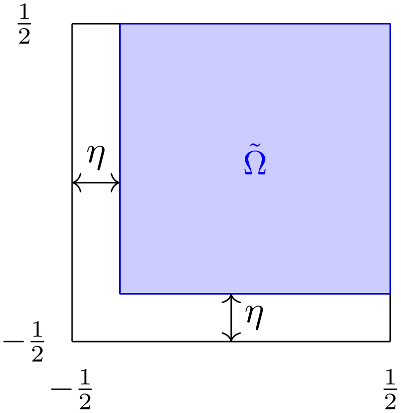


Fig. 7 Illustration of the two-dimensional extension

Let the function $g\colon \tilde{\Omega} \to \mathbb{C}$ fulfill the boundary conditions

$$g\big|_{x_i=\frac{1}{2}} \in V_n^{(d-1)}\big|_{[-\frac{1}{2}+\eta,\frac{1}{2}]}, \quad g\big|_{x_i=-\frac{1}{2}+\eta} \in V_n^{(d-1)}\big|_{[-\frac{1}{2}+\eta,\frac{1}{2}]} \textit{ for all } i \in [d],$$
$$|g^{(\alpha \boldsymbol{e}_i)}(\boldsymbol{x})| < \infty, \textit{ for all } \boldsymbol{x} \textit{ with } x_i \in \{-\tfrac{1}{2}+\eta, \tfrac{1}{2}\} \textit{ for all } \alpha = 0, \dots, m-2,$$

where $\boldsymbol{e}_i$ denotes the unit vector $(\boldsymbol{e}_i)_j = \delta_{i,j}$. Let furthermore the extension parameter fulfill

$$\eta \geqslant \frac{m-1}{2^{\left\lceil \frac{n}{d} \right\rceil + 1}}, \tag{5.4}$$

where $\lceil a \rceil$ denotes the smallest integer, which is bigger than a. Then, there exists an extension

$$\tilde{g}\colon [-\tfrac{1}{2}, -\tfrac{1}{2}+\eta]^d \to \mathbb{C}, \quad \tilde{g} \in V_n^{(d)}\big|_{[-\frac{1}{2},-\frac{1}{2}+\eta]^d}.$$

Proof Let us begin with the one-dimensional case. The function space $V_n^{(1)}$ is the space of all spline functions, which are piecewise polynomials of degree m with discontinuities of the $(m-1)$-th derivative only at the grid points $\{\frac{k}{2^{n+1}} \mid -2^n \leqslant k \leqslant 2^n\}$. These grid points divide the domain $\mathbb{T}$ naturally in pieces of length $2^{-(n+1)}$. The function g is defined on $2^{n+1} - (m-1)$ of them. The function $\tilde{g}$ has to be a piecewise polynomial of degree $m-1$ with $m-1$ pieces. To construct the coefficients of the function $\tilde{g}$, one has to solve a system of linear equations, which are independent. In fact, we have $m\,(m-1)$ coefficients and $(2+(m-2))\,(m-1) = m\,(m-1)$ constraints (at the boundary and the conditions for piecewise polynomials). This system has always been a solution. Figure 8 is an illustration of the one-dimensional case. For $m = 2$, a simple linear interpolation between $g(-\frac{1}{2})$ and $g(-\frac{1}{2}+\eta)$ does the job.

For the multivariate case, we have to observe that we use the index set I_n of hyperbolic structure. That means we need an index $\boldsymbol{j}$ with $|\boldsymbol{j}|_1 \leqslant n$, such that $\eta \geqslant \frac{m-1}{2^{(j_i+1)}}$ for all $i = 1, \dots, d$. Since all indices j_i are natural numbers, multiplication and taking the d-th root of these inequalities lead to the condition (5.4). □

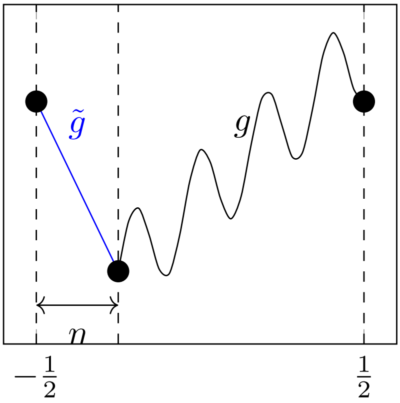


Fig. 8 The extension of the function g to $\tilde{g}$ for the dimension $d = 1$ and the order of wavelets $m = 2$

Remark 5.3 If we choose not the hyperbolic index set I_n, but the tensor index set

$$\{(\boldsymbol{j}, \boldsymbol{k}) \mid -\mathbf{1} \leqslant \boldsymbol{j} \leqslant n\,\mathbf{1}, \boldsymbol{k} \in \mathcal{I}_{\boldsymbol{j}}\},$$

the corresponding tensor product spaces are

$$V_n^{(d),\square} \colon = \bigotimes_{i=1}^{d} V_n^{(1)} = \operatorname{span}\{\psi_{\boldsymbol{j},\boldsymbol{k}}^{\mathrm{per}} \mid \|\boldsymbol{j}\|_\infty \leqslant n, \boldsymbol{k} \in \mathcal{I}_{\boldsymbol{j}}\}.$$

For the case $d = 1$, this coincides with the previous lemma. But for $d \geqslant 2$, we need in the previous proof only an index $\boldsymbol{j}$ with $|\boldsymbol{j}|_\infty \leqslant n$, such that $\eta \geqslant \frac{m-1}{2^{(j_i+1)}}$ for all $i = 1, \ldots, d$. This leads to the condition $\eta \geqslant \frac{m-1}{2^{n+1}}$.

Remark 5.4 Consider the ANOVA decomposition (3.11) of a function $f \in L_2([0,1]^d, \varrho)$. One ANOVA term $f_{\boldsymbol{u}}(\boldsymbol{y}_{\boldsymbol{u}})$ is a function, which depends on only $|\boldsymbol{u}|$ variables. Transforming $f_{\boldsymbol{u}}$ to $f_{\boldsymbol{u}} \circ \mathrm{R}_{\boldsymbol{u}}^{-1}$ needs only a $|\boldsymbol{u}|$- dimensional extension. For that reason, it is enough to choose $\eta \geqslant \frac{m-1}{2^{\left\lceil \frac{n}{|\boldsymbol{u}|} \right\rceil + 1}}$ for the transformation $\mathrm{R}_{\boldsymbol{u}}$. We will go more into details in Chapter 7.

Motivated by the previous lemma, we make the reasonable choice

$$\eta = \frac{m-1}{2^{\left\lceil \frac{n}{d} \right\rceil + 1}}. \tag{5.5}$$

In the remaining part of this section, we will focus on the one-dimensional case, because this can be also applied to high-dimensional functions with only low-dimensional interactions. The following ideas can be generalized to $d > 1$, but in this case, we have to omit boundary wavelets to ensure stability.

The one-dimensional case

The following lemma shows that the projection operator (2.6) applied to the extended function $\tilde{f}$ indeed inherits the approximation rate set by the order of the wavelets if the non-periodic function f is smooth enough on $[0,1]$. Comparing the following result with the periodic setting (2.7), the only difference is the term $\frac{1}{(1-\eta)^{3/2}}$.

Theorem 5.5 *Let $d=1$ and the maximal wavelet level $n \in \mathbb{N}$. We choose the extension parameter η as in (5.5) and the transformation (5.3), which gives $\tilde{\Omega} = [-1/2+\eta, 1/2]$. Then, we have for the approximation error of the projection operator P_n defined in (2.6) for functions $f \in H^m([0,1],\varrho)$ that*

$$\|f - (P_n\tilde{f}) \circ \tilde{\mathrm{R}}\|_{L_2([0,1],\varrho)} \lesssim \frac{2^{-nm}}{(1-\eta)^{3/2}} \|f\|_{H^m([0,1],\varrho)},$$

where $\tilde{f} = f \circ \tilde{\mathrm{R}}^{-1}$.

Proof For $j \in \mathbb{N}_0$, let us split the indices k into the sets, depending on the support of the wavelet functions

$$\begin{aligned} I_{\mathrm{in}} &= \{k \mid \operatorname{supp} \psi_{j,k}^{\mathrm{per}} \subset \tilde{\Omega}\}, \\ I_{\mathrm{bo}} &= \{k \mid \operatorname{supp} \psi_{j,k}^{\mathrm{per}} \subset [-\tfrac{1}{2}, -\tfrac{1}{2}+\eta]\}, \\ I_{\mathrm{r}} &= \mathcal{I}_j \backslash (I_{\mathrm{in}} \cup I_{\mathrm{bo}}). \end{aligned}$$

First, we have a look at the wavelet coefficients of the extended function $\tilde{f}$ with $j > n$. From [26, Lemma 3.4], we have that

$$\langle \tilde{f}, \psi_{j,k}^{\mathrm{per}} \rangle = 2^{j/2} 2^{-jm} \int_{I_{j,k}} \overline{\tilde{f}^{(m)}}(x) \Psi_m(2^j x - k)\, \mathrm{d}x, \tag{5.6}$$

where $I_{j,k} = \operatorname{supp} \psi_{j,k}^{\mathrm{per}} \subset \mathbb{T}$ and the function Ψ_m is defined in (B.2). Since the extension of f to $\tilde{f}$, namely $\tilde{f}\,|_{\mathbb{T}\backslash\tilde{\Omega}}$, is contained in the space of wavelet functions, the m-th derivative of $\tilde{f}$ is zero at the boundary $(-1/2, -1/2+\eta)$, and we have that

$$\tilde{f}^{(m)}(x) = \delta(x+\tfrac{1}{2}) F_1(f) + \delta(x+\tfrac{1}{2}-\eta) F_0(f), \quad x \in [-\tfrac{1}{2}, -\tfrac{1}{2}+\eta],$$

where $\delta(\cdot)$ is the delta distribution and the numbers $F_0(f)$ and $F_1(f)$ depend on the boundary behavior of the function f.
For the indices $k \in \mathcal{I}_{\mathrm{r}}$, we split the integral (5.6),

$$\langle \tilde{f}, \psi_{j,k}^{\mathrm{per}} \rangle = 2^{j/2} 2^{-jm} \left(\int_{I_{j,k} \cap \tilde{\Omega}} \overline{\tilde{f}^{(m)}}(x) \Psi_m(2^j x - k)\, \mathrm{d}x + F_1(f) \Psi_m(2^j(-\tfrac{1}{2}+\eta) - k) \right), \tag{5.7}$$

if $-1/2+\eta \in \operatorname{supp} \psi_{j,k}^{\mathrm{per}}$. The other case where $-1/2 \in \operatorname{supp} \psi_{j,k}^{\mathrm{per}}$ is analogue. The function Ψ_m has the property that $\Psi_m(x) = 0$ for $x \in \mathbb{N}$ (see Lemma B.2). Because of the choice $\eta = \frac{m-1}{2^{n+1}}$ and the assumption $j > n$, the numbers $2^j(-\frac{1}{2} + \eta) - k$ are in $\{1, \ldots, m\}$. This allows us to omit the second term in (5.7), and we get that

$$\langle \tilde{f}, \psi_{j,k}^{\mathrm{per}} \rangle = \begin{cases} 2^{j/2} 2^{-jm} \int_{I_{j,k} \cap \tilde{\Omega}} \overline{\tilde{f}^{(m)}}(x) \Psi_m(2^j x - k) \, \mathrm{d}x, & \text{if } k \in I_{\mathrm{r}} \cup I_{\mathrm{in}}, \\ 0 & \text{if } k \in I_{\mathrm{bo}}. \end{cases}$$

Using Cauchy-Schwarz inequality, we receive for $k \in I_{\mathrm{r}} \cup I_{\mathrm{in}}$ that

$$\begin{aligned} |\langle \tilde{f}, \psi_{j,k}^{\mathrm{per}} \rangle| &\leqslant 2^{j/2} 2^{-jm} \left(\int_{I_{j,k} \cap \tilde{\Omega}} |\tilde{f}^{(m)}(x)|^2 \mathrm{d}x \right)^{1/2} \left(\int_{I_{j,k} \cap \tilde{\Omega}} |\Psi_m(2^j x - k)|^2 \mathrm{d}x \right)^{1/2} \\ &\lesssim 2^{-jm} \, \|\tilde{f}^{(m)}\|_{L_2(I_{j,k} \cap \tilde{\Omega})} \max_{x \in [0, 2m-1]} \Psi_m(x) \\ &\lesssim 2^{-jm} \frac{1}{(1-\eta)^{1/2}} \, \|f\|_{H^m([0,1],\varrho)} \Big|_{\mathrm{R}^{-1}(I_{j,k} \cap \tilde{\Omega})}, \end{aligned} \tag{5.8}$$

where the last inequality follows from (4.19) and (B.3). The Riesz basis property (2.4), which also applies to the dual wavelets yields

$$\sum_{k \in \mathcal{I}_j} |\langle f, \psi_{j,k}^{\mathrm{per}*} \rangle|^2 \lesssim \frac{1}{\gamma_m} \left\| \sum_{k \in \mathcal{I}_j} \langle f, \psi_{j,k}^{\mathrm{per}*} \rangle \psi_{j,\boldsymbol{k}}^{\mathrm{per}} \right\|_{L_2(\mathbb{T})}^2 = \frac{1}{\gamma_m} \left\| \sum_{k \in \mathcal{I}_j} \langle f, \psi_{j,k}^{\mathrm{per}} \rangle \psi_{j,k}^{\mathrm{per}*} \right\|_{L_2(\mathbb{T})}^2 \lesssim \frac{\delta_m}{\gamma_m} \sum_{k \in \mathcal{I}_j} |\langle f, \psi_{j,k}^{\mathrm{per}} \rangle|^2. \tag{5.9}$$

Also, because of the Riesz basis property of the wavelet functions, we have for $j > n$ that

$$\int_{-1/2+\eta}^{1/2} \left| \sum_{k=0}^{2^j-1} a_{j,k} \psi_{j,k}^{\mathrm{per}}(x) \right|^2 \mathrm{d}x \leqslant \int_{-1/2}^{1/2} \left| \sum_{k \in I_{\mathrm{in}} \cup I_{\mathrm{r}}} a_{j,k} \psi_{j,k}^{\mathrm{per}}(x) \right|^2 \mathrm{d}x \leqslant \delta_m \sum_{k \in I_{\mathrm{in}} \cup I_{\mathrm{r}}} |a_{j,k}|^2. \tag{5.10}$$

To estimate the error of the projection operator P_n, we have to estimate the sum of wavelet coefficients, namely we first insert the definition (2.6) of P_n,

$$\begin{aligned} \|f - (P_n \tilde{f}) \circ \mathrm{R}\|_{L_2([0,1],\varrho)}^2 &= \frac{1}{1-\eta} \, \|\tilde{f} - P_n \tilde{f}\|_{L_2(\tilde{\Omega})}^2 = \frac{1}{1-\eta} \, \| \sum_{|j|>n} \sum_{k \in \mathcal{I}_j} \langle \tilde{f}, \psi_{j,k}^{\mathrm{per},*} \rangle \psi_{j,k}^{\mathrm{per}} \|_{L_2(\tilde{\Omega})}^2 \\ &\leqslant \frac{1}{1-\eta} \left(\sum_{|j|>n} \| \sum_{k \in \mathcal{I}_j} \langle \tilde{f}, \psi_{j,k}^{\mathrm{per},*} \rangle \psi_{j,k}^{\mathrm{per}} \|_{L_2(\tilde{\Omega})} \right)^2 \\ &\overset{(5.10)}{\leqslant} \frac{\delta_m^{1/2}}{1-\eta} \left(\sum_{|j|>n} \left(\sum_{k \in \mathcal{I}_j} |\langle \tilde{f}, \psi_{j,k}^{\mathrm{per},*} \rangle|^2 \right)^{1/2} \right)^2 \\ &\overset{(5.9)}{\leqslant} \frac{\delta_m^{3/2}}{\gamma_m \, (1-\eta)} \left(\sum_{|j|>n} \left(\sum_{k \in \mathcal{I}_j} |\langle \tilde{f}, \psi_{j,k}^{\mathrm{per}} \rangle|^2 \right)^{1/2} \right)^2 \end{aligned}$$

$$\begin{aligned}
&\overset{(5.8)}{\lesssim} \frac{\delta_m^{3/2}}{\gamma_m\,(1-\eta)^3} \left(\sum_{|j|>n} 2^{-jm} \left(\sum_{k\in\mathcal{I}_j} \|f\|^2_{H^m([0,1],\varrho)|_{\mathrm{R}^{-1}(I_{j,k}\cap\tilde{\Omega})}} \right)^{1/2} \right)^2 \\
&\lesssim \frac{\delta_m^{3/2}}{\gamma_m\,(1-\eta)^3} \|f\|^2_{H-([0,1],\varrho)} \left(\sum_{|j|>n} 2^{-jm} \right)^2 . \qquad (5.11)
\end{aligned}$$

The last sum of 2^{-jm} is bounded by geometric series,

$$\sum_{|j|>n} 2^{-jm} \leqslant \frac{1}{1-2^{-m}} - \frac{1-2^{-(n+1)m}}{1-2^{-m}} = \frac{2^{-(n+1)m}}{1-2^{-m}} \lesssim 2^{-nm}.$$

Taking the square root in (5.11) gives the result, where the factor $\frac{\delta_m^{3/2}}{\gamma_m}$ is a constant.□

Note that we receive at least in the one-dimensional case the same approximation rate as in the periodic setting. In the higher-dimensional setting, this is not the case, since we lose the orthogonality between different wavelet levels in the $L_2(\tilde{\Omega})$-norm. It is also possible to estimate the L_∞-error. Here, also, the only change is the additional term $\frac{1}{1-\eta}$ in comparison to the periodic result, (2.8).

Theorem 5.6 *Let $d = 1$ and the maximal wavelet level $n \in \mathbb{N}$. We choose the extension parameter η as (5.5), the transformation (5.3) and denote $\tilde{f} = f \circ \mathrm{R}^{-1}$. Then, we have for the approximation error of the projection operator*

$$\|f - (P_n\tilde{f}) \circ \mathrm{R}\|_{L_\infty([0,1])} \lesssim \frac{2^{-n(m-1/2)}}{1-\eta} \|f\|_{H^m([0,1],\varrho)} \, .$$

Proof Similar to the proof of the previous theorem, we consider

$$\begin{aligned}
\|f - (P_n\tilde{f}) \circ \mathrm{R}\|_{L_\infty([0,1])} &= \sup_{y\in[0,1]} |f(y) - (P_n\tilde{f}) \circ \mathrm{R}(y)| = \sup_{x\in[-\frac{1}{2}+\eta,\frac{1}{2}]} |\tilde{f}(x) - (P_n\tilde{f})(x)| \\
&= \sup_{x\in[-\frac{1}{2}+\eta,\frac{1}{2}]} |\sum_{|j|>n} \sum_{k\in\mathcal{I}_j} \langle \tilde{f}, \psi_{j,k}^{\mathrm{per},*} \rangle \psi_{j,k}^{\mathrm{per}}(x)|.
\end{aligned}$$

Using the same lines as in [26, Theorem 3.15], we have

$$\begin{aligned}
\|f - (P_n\tilde{f}) \circ \mathrm{R}\|_{L_\infty([0,1])} &\lesssim \left(\sum_{|j|>n} 2^{-|j|_1(m-1/2)} \right) \left(\sup_{|j|>n} 2^{|j|_s} \left(\sum_{k\in\mathcal{I}_j} |\langle \tilde{f}, \psi_{j,k}^{*,\mathrm{per}} \rangle|^2 \right)^{1/2} \right) \\
&\lesssim \frac{2^{-n(m-1/2)}}{1-\eta} \|f\|_{H^m([0,1],\varrho)} \, ,
\end{aligned}$$

which gives the assertion.

In the following, we discuss the numerical properties, that arise when using such an extension. We lose some stability in the sense that the wavelet matrix $\boldsymbol{A}$ has a bigger condition number. But nevertheless, we estimate the eigenvalues of the Moore Penrose inverse from below. The eigenvalues of the expectation matrix

$$\mathbf{\Lambda} = \left(\int_{\mathbb{T}} \psi_{j,k}^{\mathrm{per}}(x)\, \psi_{j',k'}^{\mathrm{per}}(x)\, \mathrm{d}x \right)_{(j,k)\in I_n,(j',k')\in I_n}$$

are bounded by the Riesz constants γ_m and δ_m (see [26, Lemma 3.18]). The transformation R changes the expectation matrix to

$$\tilde{\mathbf{\Lambda}} = \left(\int_{\tilde{\Omega}} \psi_{j,k}^{\mathrm{per}}(x)\, \psi_{j',k'}^{\mathrm{per}}(x)\, \mathrm{d}x \right)_{(j,k)\in I_n,(j',k')\in I_n}.$$

If we choose the extension parameter η like (5.5), it turns out that the eigenvalues of $\tilde{\mathbf{\Lambda}}$ do not differ much from the eigenvalues of the initial expectation matrix $\mathbf{\Lambda}$. We show this numerically. We lose the orthogonality of the wavelets of different levels. But the entries of the matrix $\tilde{\mathbf{\Lambda}}$ differ from the entries of $\mathbf{\Lambda}$ only at indices where $\operatorname{supp} \psi_{j,k}^{\mathrm{per}} \cap [-\frac{1}{2}, -\frac{1}{2}+\eta] \neq \emptyset$. For different maximal level n, we construct the matrix $\tilde{\mathbf{\Lambda}}$ and calculate the extremal eigenvalues $\mu_{\min}(\tilde{\mathbf{\Lambda}})$ and $\mu_{\max}(\tilde{\mathbf{\Lambda}})$. The results are summarized in Table 1. For comparison, we also give the extremal eigenvalues of the matrix $\mathbf{\Lambda}$, which are the lower Riesz-constant γ_m and 1. We leave the proof that $\mu_{\min}(\tilde{\mathbf{\Lambda}}) > C > 0$ for a constant C is an open problem, but the numeric indicates that this is true.

In fact, the choice of η must be a balance between an ill-conditioned matrix for big η and a large approximation error for small η. The choice (5.5) does this job.

Error estimates

Up to now, we gave estimates for the L_2 and L_∞ error of the projection operator P_n, (2.6), instead of the approximation operator $S_n^{\mathcal{X}}$. To end this subsection, we give also for the non-periodic case an error estimate with high probability, similar to Theorem 5.1.

Corollary 5.7 *Let $d = 1$, $m \in \mathbb{N}$ be the order of vanishing moments of the wavelet ψ, $\varrho \in C^{m-1}([0,1])$ be a density and and $n \in \mathbb{N}$ be the maximal level of the wavelets. We choose the transformation* R *as* (5.3) *with the extension parameter η as in* (5.5). *Let*

Table 1 Extremal eigenvalues of the expectation matrix $\tilde{\mathbf{\Lambda}}$ for different maximal level n and order m of the wavelets

m	n	2	3	4	5	6	7	$\mathbf{\Lambda}$
2	$\mu_{\min}(\tilde{\mathbf{\Lambda}})$	0.0896	0.0903	0.0879	0.0859	0.0848	0.0843	0.1481
	$\mu_{\max}(\tilde{\mathbf{\Lambda}})$	0.8990	0.9497	0.9748	0.9873	0.9937	0.9968	1
3	$\mu_{\min}(\tilde{\mathbf{\Lambda}})$	0.0032	0.0025	0.0024	0.0024	0.0024	0.0025	0.0379
	$\mu_{\max}(\tilde{\mathbf{\Lambda}})$	0.7735	0.8854	0.9426	0.9714	0.9857	0.9928	1

furthermore $\mathcal{Y} = (y_i)_{i=1}^M \subset [0, 1]$ *with* $M \gtrsim r\, 2^{n+1}\, (n+1)$ *be drawn i.i.d. at random according to* ϱ *and* $r > 1$. *Then, we denote by* $\mathcal{X} = \mathrm{R}(\mathcal{Y})$ *the transformed samples. Let furthermore* $\mu_{\min}(\tilde{\mathbf{\Lambda}}) \geqslant C > 0$. *Then, we have*

$$\mathbb{P}\left(\|f - (S_n^{\mathcal{X}} \tilde{f}) \circ \mathrm{R}\|_{L_2([0,1],\varrho)} \lesssim \frac{2^{-nm}}{(1-\eta)^{5/2}} \, \|f\|_{H^m([0,1],\varrho)} \right) \geqslant 1 - 2M^{-r}.$$

Proof First, we denote $e_2 := \|\tilde{f} - (P_n \tilde{f})\|_{L_2(\tilde{\Omega})}$ and $e_\infty := \|\tilde{f} - (P_n \tilde{f})\|_{L_\infty(\tilde{\Omega})}$. By using the extension, we lose the orthogonality of the wavelet functions even for different levels, since the L_2-norm is then defined on $\tilde{\Omega}$ and not on the whole torus. Therefore, we have to modify the proof of [26, Theorem 3.20] slightly. We have

$$\begin{aligned} \|f - (S_n^{\mathcal{X}} \tilde{f}) \circ \mathrm{R}\|_{L_2([0,1],\varrho)} &= \frac{1}{1-\eta} \, \|\tilde{f} - S_n^{\mathcal{X}} \tilde{f}\|_{L_2(\tilde{\Omega})} \leqslant \frac{1}{1-\eta} \left(e_2 + \|P_n \tilde{f} - S_n^{\mathcal{X}} \tilde{f}\|_{L_2(\tilde{\Omega})} \right) \\ &= \frac{1}{1-\eta} \left(e_2 + \|S_n^{\mathcal{X}} (P_n \tilde{f} - \tilde{f})\|_{L_2(\mathbb{T}^d)} \right) \leqslant \frac{1}{1-\eta} \left(e_2 + \|S_n^{\mathcal{X}}\|_2 \, \|P_n f - f\|_{\ell_2(\mathcal{X})} \right). \end{aligned}$$

For the last term, we use the same lines as in [26, Theorem 3.20], which are based on Bernstein's inequality to get

$$\mathbb{P}\left(\|\tilde{f}(x_i) - P_n \tilde{f}(x_i)\|_{\ell_2(\mathcal{X})} \geqslant \left(M \sqrt{\frac{2\, e_\infty^2 e_2^2 r \log M}{M}} + \frac{2\, e_\infty^2 r \log M}{3M} + e_2^2 \right)^{1/2} \right) \leqslant M^{-r}.$$

Taking the event into account that

$$\|S_n^{\mathcal{X}}\|_2 \leqslant \sqrt{\frac{2}{M \mu_{\min}(\tilde{\mathbf{\Lambda}})}} < \sqrt{\frac{2}{C}}$$

with high probability, we obtain by union bound the overall probability exceeding the sum of the probabilities, i.e.,

$$\mathbb{P}\left(\|\tilde{f} - S_n^{\mathcal{X}} \tilde{f}\|_{L_2(\tilde{\Omega})} \leqslant e_2 + \sqrt{\tfrac{2}{C}} \left(\tfrac{e_2^2}{M} + e_2 e_\infty \sqrt{\tfrac{r \log M}{M}} + e_\infty^2 \tfrac{r \log M}{M} \right)^{1/2} \right) \geqslant 1 - 2\, M^{-r}.$$

Collecting the bounds from the occurring terms from Theorems 5.5 and 5.6 as well as logarithmic oversampling, which means $\frac{\log M}{M} \lesssim 2^{-n}$, we end up with

$$\begin{aligned} &\|f - (S_n^{\mathcal{X}} \tilde{f}) \circ \mathrm{R}\|_{L_2([0,1],\varrho)} = \frac{1}{1-\eta} \, \|\tilde{f} - S_n^{\mathcal{X}} \tilde{f}\|_{L_2(\tilde{\Omega})} \\ &\lesssim \frac{1}{1-\eta} \left(\frac{2^{-nm}}{(1-\eta)^{3/2}} + \sqrt{\tfrac{2}{C}} \left(\frac{2^{-2nm}}{(1-\eta)^3\, 2^{n+1}} + \frac{2^{-2nm}}{(1-\eta)^{5/2}} \sqrt{r} + \frac{2^{-2nm}}{(1-\eta)^2}\, r \right)^{1/2} \right) \cdot \|f\|_{H^m([0,1],\varrho)} \\ &= \frac{2^{-nm}}{(1-\eta)^{5/2}} \left(1 + \sqrt{\tfrac{2}{C}} \left(\frac{1}{(1-\eta)^{3/2}\, 2^{n+1}} + \frac{\sqrt{r}}{(1-\eta)} + \frac{r}{(1-\eta)^{1/2}} \right)^{1/2} \right) \cdot \|f\|_{H^m([0,1],\varrho)} \end{aligned}$$

with high probability. □

5.2 Numerical experiments

In this section, we study the approximation behavior numerically for some examples to underpin our findings from Theorem 5.1 and Corollary 5.7. We consider examples where $\Omega \in \{\mathbb{R}^d, \mathbb{T}^d, [0,1]^d\}$.
We do the following procedure. For a maximal level n, we draw $M \gtrsim N \log N \asymp 2^n\, n^d$ i.i.d. samples according to the density ϱ, which coincides with logarithmic oversampling. Also, the corresponding function values are given. In our approximation, we transfer the samples in the set $\mathcal{Y}$ to the torus, by $\mathrm{R}(\mathcal{Y}) = \mathcal{X}$ and apply the approximation operator $S_n^{\mathcal{X}}$ given in (2.3). A good estimator for the $L_2(\Omega, \varrho)$-error, is the *root mean square error* (RMSE), which is defined by

$$\mathrm{RMSE} = \left(\frac{1}{|\mathcal{Y}_{\text{test}}|} \sum_{\boldsymbol{y} \in \mathcal{Y}_{\text{test}}} |f(\boldsymbol{y}) - (S_n^{\mathcal{X}}(f \circ \mathrm{R}^{-1})) \circ \mathrm{R}(\boldsymbol{y})|^2 \right)^{1/2}, \tag{5.12}$$

for sample points $\mathcal{Y}_{\text{test}} \subset \Omega$, which are i.i.d. according to ϱ. We use always $|\mathcal{Y}_{\text{test}}| = 3\,|\mathcal{Y}| = 3\,M$. We defined in Examples 4.3 and 4.11 only one-dimensional densities. In the following, we interpret the d-dimensional densities as a tensor product in the sense of (3.7).

Distributions on $\mathbb{R}^d$
We begin with the normal distribution ϱ_N from (4.5) for all one-dimensional densities $\varrho_i(y_i)$. As a test function, we use the Gaussian

$$f : \mathbb{R}^d \to \mathbb{R}, \quad f(\boldsymbol{y}) = \mathrm{e}^{-\|\boldsymbol{y}\|_2^2}, \tag{5.13}$$

which is smooth. But in the sense of (4.10), we have to note that in the one-dimensional-case

$$\begin{aligned}
\|f\|^2_{H^1(\mathbb{R},\varrho_N)} &= \tfrac{1}{\sqrt{5}} + \tfrac{8\pi}{3\sqrt{3}} \approx 5.28 < \infty, \\
\|f\|^2_{H^2(\mathbb{R},\varrho_N)} &= \tfrac{1}{\sqrt{5}} + (\tfrac{8\pi}{3\sqrt{3}} + 48\pi^2) + 144\pi^2 \approx 1900 < \infty, \\
\|f\|^2_{H^3(\mathbb{R},\varrho_N)} &= \infty.
\end{aligned}$$

Due to the embedding $H^2(\mathbb{T}) \subset \boldsymbol{B}^2_{2,\infty}(\mathbb{T})$, we also have $f \circ \mathrm{R}^{-1} \in \boldsymbol{B}^2_{2,\infty}(\mathbb{T})$. To investigate the smoothness further, we have a look at the terms from Definition 4.9, i.e., for $s = {}^5\!/_2$

$$\sup_{\substack{j \in \mathbb{Z} \\ j \neq 0}} 2^{5|j|_1} \sup_{\alpha \leqslant 2} \sup_{1 \leqslant \beta \leqslant \alpha} \sum_{k \in J_j} \frac{1}{|k|^{2\alpha}} \left| c_k^{\varrho} \left(\mathrm{D}^{\beta} f(y) B_{\alpha,\beta}(y) \right) \right|^2 .$$

We have

$$\begin{aligned}
|c_k^{\varrho}(\mathrm{D}^2 f(y) B_{2,2}(y))| &= |c_k^{\varrho}((-2+4y^2)\mathrm{e}^{-y^2}\, 2\pi\, \mathrm{e}^{y^2})| = |c_k^{\varrho}((-2+4y^2)\, 2\pi)| \\
&= |2\pi\, c_k(-2+4(\mathrm{R}^{-1}(x))^2)| \asymp \tfrac{1}{|k|},
\end{aligned}$$

since $\mathrm{R}^{-1}(x)^2$ is a smooth function except at $x = 1/2$. Analogously, we get for the other term

$$|c_k^{\varrho}(\mathrm{D}^1 f(y) B_{2,1}(y))| \asymp \frac{1}{|k|}.$$

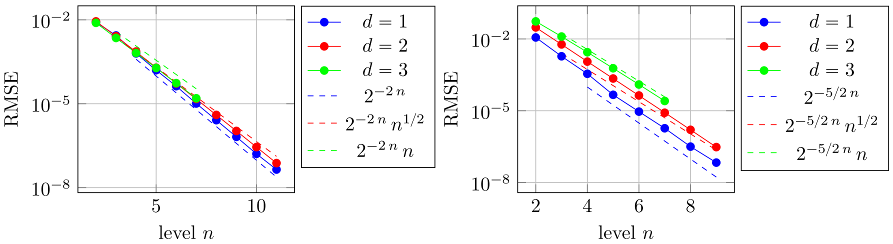
This yields

$$\|f\|^2_{\boldsymbol{B}^{5/2}_{2,\infty}} \lesssim \sup_{j\in\mathbb{Z}} 2^{5|j|} \sum_{k\in J_j} \frac{1}{|k|^4}\frac{1}{|k|^2} \lesssim \sup_{j\in\mathbb{Z}} 2^{5|j|}2^{|j|-1}2^{-6|j|} < \infty,$$

since the index sets J_j are defined such that $k > 2^{|j|-1}$ and $|J_j| = 2^{|j|}$. The function f and the density ϱ_N have a tensor product structure; hence, we get $f \circ \mathrm{R}^{-1} \in \boldsymbol{B}^{5/2}_{2,\infty}(\mathbb{T}^d)$ and $f \in \boldsymbol{B}^{5/2}_{2,\infty}(\Omega, \varrho_N)$.

We did the approximation for $d \in \{1, 2, 3\}$ and for the order $m = \{2, 3\}$ of vanishing moments of the wavelet. In Fig. 9, we plotted the results. One can see that we end up with the proposed error decay rates from Theorem 5.1. In case where $m = 2$, we are in the case (5.1) and get the proposed error decay rate of $2^{-2n}\, n^{(d-1)/2}$. If we increase the order of vanishing moments of the wavelets to $m = 3$, we are in the case (5.1) and receive also numerically the proposed error decay of $2^{-5/2n}\, n^{(d-1)/2}$. The numerical results are even slightly better in some cases.

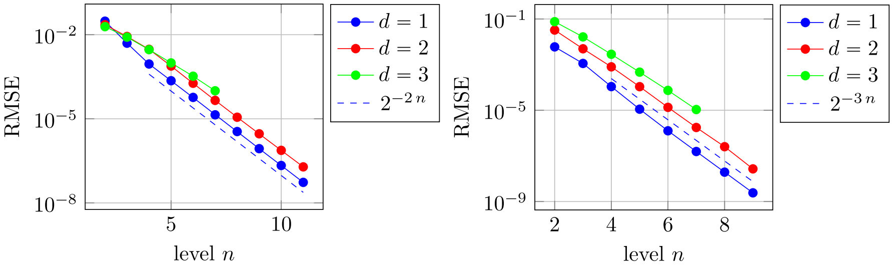
Note that a different density $\varrho(y)$ can lead to a different smoothness of $f \circ \mathrm{R}^{-1}$ even for the same function f. For a tensor product of the one-dimensional Cauchy ϱ_C, (4.6), or the Laplace distribution ϱ_L, (4.7), we even have $f \circ \mathrm{R}^{-1} \in H^s_{\mathrm{mix}}(\Omega, \varrho)$ for all $s \in \mathbb{N}$. That follows from $\mathrm{D}^{\alpha} f(y) = p(y)\mathrm{e}^{-y^2/2}$ for a polynomial $p(y)$ and the fact that all differentials of the Cauchy as well as the Laplace distribution are polynomials or polynomials with a the factor of the behavior $\mathrm{e}^{k|y-2|}$. Furthermore, all integrals of the form $\int_{\mathbb{R}} \mathrm{e}^{-y^2+k|y-2|} p(y)\,\mathrm{d}y$ are finite. In Figs. 10 and 11, we plotted the approximation results. In this case, the order of vanishing moments determines the error decay. In dimension $d = 3$ with Laplace distribution, we are still in the preasymptotic case.



(a) Decay of the RMSE for $m = 2$.

(b) Decay of the RMSE for $m = 3$.

Fig. 9 Approximation of the function (5.13) on $\mathbb{R}^d$ for $d \in \{1, 2, 3\}$ and the normal distribution ϱ_N



(a) Decay of the RMSE for $m = 2$.

(b) Decay of the RMSE for $m = 3$.

Fig. 10 Approximation of the function (5.13) on $\mathbb{R}^d$ for $d \in \{1, 2, 3\}$ with the Cauchy distribution ϱ_C

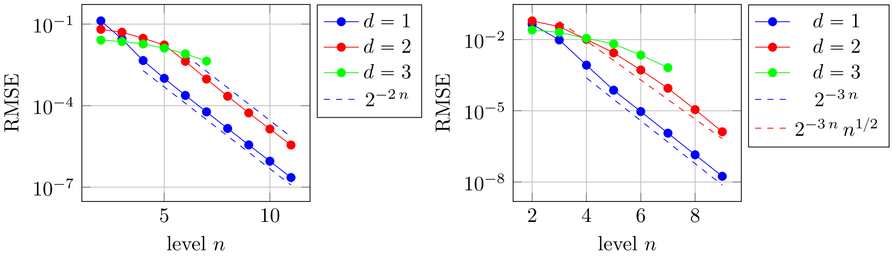
The beta distribution on the torus

Let us consider the beta distribution (4.20), but shifted by $-1/2$ to the torus $\mathbb{T} = [-1/2, 1/2)$, for all one-dimensional densities $\varrho_i(x_i)$, which also includes the uniformly distribution on $\mathbb{T}$ for $\alpha = 1$. We choose as test function

$$f : \mathbb{T}^d \to \mathbb{R}, \quad f(\boldsymbol{y}) = \prod_{i=1}^{d} (y_i - \tfrac{1}{2})^3 (y_i + \tfrac{1}{2})^3, \tag{5.14}$$

which is the tensor product of a polynomial of degree 6 and has triple zeros at $1/2$ and is in $H^3_{\text{mix}}(\mathbb{T}^d)$. Depending on the choice of the parameter α of the beta distribution, the transformed function has different regularity. We consider first the one-dimensional case. The crucial points to decide whether the norm $\|f\|_{L_2(\mathbb{T},\varrho_{B,\alpha})}$ is finite is at the points with lower regularity $1/2 \cong -1/2$. Because of the symmetry of the function f as well as the density ϱ, it is sufficient to have a look at the point $y = 1/2$. There we have the behavior $f^{(i)} \sim (y - 1/2)^{3-i}$ for $i = 0, 1, 2, 3$. With the same arguments we have that $\varrho^{(i)}_{B,\alpha} \sim (y - 1/2)^{\alpha-1-i}$ for $\alpha \neq 1, 2$ and $i = 0, 1, 2$. Furthermore, the integral $\int_0^1 x^k \,\mathrm{d}x$ is finite for $k \in \mathbb{R}$ and $k > -1$. Hence, Definition 4.1 gives that we have to require the following:

$$f \in H^1(\mathbb{T}, \varrho_{B,\alpha}) \Rightarrow f' \in L_2(\mathbb{T}, \Upsilon_{1,1}) \qquad \Rightarrow \alpha < 6,$$
$$f \in H^2(\mathbb{T}, \varrho_{B,\alpha}) \Rightarrow f'' \in L_2(\mathbb{T}, \Upsilon_{2,2}),\ f' \in L_2([0, 1], \Upsilon_{1,2}) \qquad \Rightarrow \alpha < 2,$$



(a) Decay of the RMSE for $m = 2$.

(b) Decay of the RMSE for $m = 3$.

Fig. 11 Approximation of the function (5.13) on $\mathbb{R}^d$ for $d \in \{1, 2, 3\}$ with the Laplace distribution ϱ_L

$$f \in H^3(\mathbb{T}, \varrho_{B,\alpha}) \Rightarrow f''' \in L_2(\mathbb{T}, \Upsilon_{3,3}),\ f'' \in L_2([0,1], \Upsilon_{2,3}),\ f' \in L_2([0,1], \Upsilon_{1,3}) \qquad \Rightarrow \alpha < 6/5.$$

Since the function f is a tensor product, we have the same estimates for the multivariate cases. Indeed, we used the order of vanishing moments $m = 3$ of the wavelets, which limits the maximal error decay rate, and Fig. 12 is the resulting numerical approximation decay for different parameters α and d. Indeed, if we are below the critical values $6/5$, 2 and 6 for α, we get the desired approximation rates of 2^{-3n}, 2^{-2n} and 2^{-n} given in Theorem 5.1.
Even dense samples at the boundary, which coincide with small α, can not increase the error decay rate of 3.

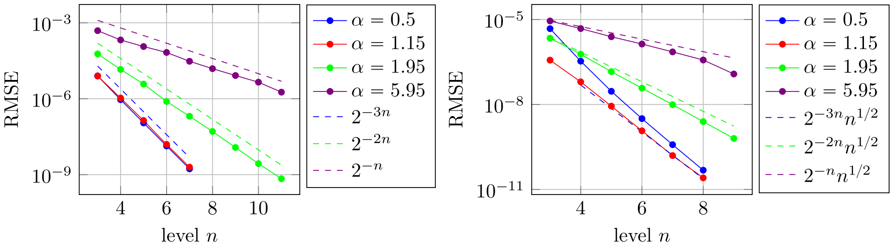
Extensions of non-periodic functions on the cube
Here, we want to demonstrate the benefits of the extension proposed in Sect. 5.1. Let us study the non-periodic function

$$f\colon [0,1] \to \mathbb{R}, \quad f(y) = y^3. \tag{5.15}$$

and the uniformly distribution on the cube, $\varrho(y) = 1$. Also, for this non-periodic function, we managed to use the periodic approximation operator and get good approximation results. We use a polynomial of degree 3, since a lower degree together with the order of the wavelets m leads to a function $f \circ \mathrm{R}^{-1}$, which is in the finite function space, which we use for the approximation and gives us approximation errors near machine precision. The results are plotted in Fig. 13. We see that the extension increases the approximation rate to 2^{-mn}, as proposed in Crollary 5.7 Additionally, the wavelet matrix is in both cases well-conditioned.

6 Second setting: using kernel density estimation for unknown density ρ

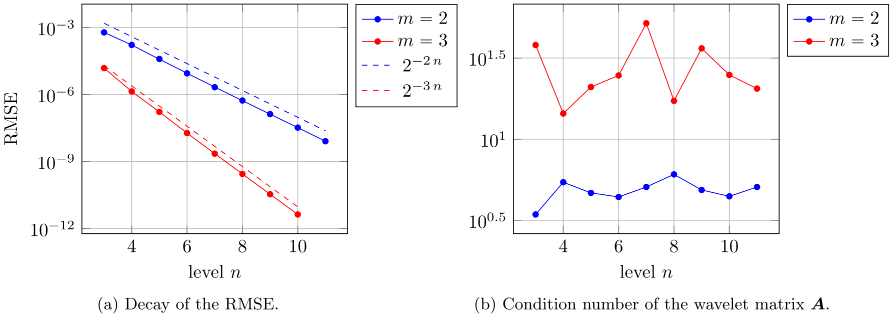
While working with real-world data, the underlying density $\varrho(\boldsymbol{y})$ is possibly a priori not known, and we only have the given random sample points $\mathcal{Y}$. For that reason, we want to adapt our strategy to this setting. A transformation of the given data $\mathcal{Y}$ is also in this case a useful tool to approximate a function f well. Instead of



(a) Decay of the RMSE for $d = 1$.

(b) Decay of the RMSE for $d = 2$.

Fig. 12 Approximation on $\mathbb{T}^d$ for $d \in \{1, 2\}$ of the test function (5.14) samples distributed with respect to the beta distribution $\varrho_{B,\alpha}$



(a) Decay of the RMSE.

(b) Condition number of the wavelet matrix $\boldsymbol{A}$.

Fig. 13 Approximation on [0, 1] of the test function (5.15) with uniformly distributed points using Chui-Wang wavelets of order $m \in \{2, 3\}$

using the transformation R belonging to the underlying density ϱ as in Chapter 5, we approximate the underlying density function by a kernel density estimation [15]. The cumulative distribution function $\mathring{\mathrm{R}}$ of the estimated density function gives us a transformation for the underlying data set $\mathcal{Y}$ to the samples $\mathring{\mathcal{X}} = \mathring{\mathrm{R}}(\mathcal{Y})$ on the torus. Then, we apply our approximation method for functions on the torus and at the end we transform the function back to a function defined on Ω. We will restrict our study in this chapter to the one-dimensional case, which we apply in Chapter 7 to high-dimensional functions. The error between the estimated density and the true density influences the total approximation error.

Let us introduce the kernel density estimator

$$\mathring{\varrho}(y) = \sum_{x \in \mathcal{X}} \frac{1}{\sigma M} k\left(\frac{y - x}{\sigma}\right), \tag{6.1}$$

where k is a non-negative kernel function $k : \Omega \to \mathbb{R}$ which is normed by $\int_\Omega k(y)\,\mathrm{d}y = 1$ and σ is a smoothing parameter. Frequently used kernels are the standard normal distribution ϱ_N, (4.5) or B-Splines (B.1). The normalization ensures that also $\mathring{\varrho}$ is normalized. We get a transformation $\mathring{\mathrm{R}} : \Omega \to \mathbb{T}$, which fulfills (3.1), by integration,

$$\begin{aligned}\mathring{\mathrm{R}}(y) &= \begin{cases} \int_0^y \mathring{\varrho}(t)\,\mathrm{d}t - \frac{1}{2} & \text{if } \Omega = [0, 1], \\ \int_{-\infty}^y \mathring{\varrho}(t)\,\mathrm{d}t - \frac{1}{2} & \text{if } \Omega = \mathbb{R}, \\ \int_{-1/2}^y \mathring{\varrho}(t)\,\mathrm{d}t - \frac{1}{2} & \text{if } \Omega = \mathbb{T}. \end{cases} \\ &= \frac{1}{M} \sum_{x \in \mathcal{X}} K(y - x) - \tfrac{1}{2}, \end{aligned} \tag{6.2}$$

where K is the antiderivative

$$K(y) = \begin{cases} \int_0^y k(t)\,\mathrm{d}t & \text{if } \Omega = [0, 1], \\ \int_{-\infty}^y k(t)\,\mathrm{d}t & \text{if } \Omega = \mathbb{R}, \\ \int_{-1/2}^y k(t)\,\mathrm{d}t & \text{if } \Omega = \mathbb{T}. \end{cases}$$

Hence, the integral K of the kernel k has to be calculated once in advance. Then, the calculation of the integral of $\mathring{\varrho}$ is only the evaluation of K at different points. Using this transformation, we get our transformed sample points by $\mathring{\mathcal{X}} = \mathring{\mathrm{R}}(\mathcal{Y})$. If $\mathring{\varrho}$ is a good approximation to ϱ, these samples $\mathring{\mathcal{X}}$ are nearly uniformly distributed on $\mathbb{T}$, which allows us to approximate the function $f \circ \mathring{\mathrm{R}}^{-1}$ using an approximation operator on $\mathbb{T}^d$, for instance $S_n^{\mathring{\mathcal{X}}}$ from (2.3). Our procedure to get an approximation of f out of the samples $\mathcal{X}$ and the corresponding function values $\boldsymbol{f}$ is summarized in Fig. 14. For shortening notation, we denote the error function $\mathring{e}_f \colon \Omega \to \mathbb{R}$ by

$$\mathring{e}_f = f - \left(S_n^{\mathring{\mathcal{X}}}(f \circ \mathring{\mathrm{R}}^{-1})\right) \circ \mathring{\mathrm{R}},$$

which we aim to estimate. Using the theory of the previous section, we receive a bound for $\|\mathring{e}_f\|_{L_2(\Omega,\mathring{\varrho})}$ in the norm with density $\mathring{\varrho}$, if we assume that we choose the bandwidth σ so that $\varrho \approx \mathring{\varrho}$, which yields that the samples $\mathring{\mathcal{X}} = \mathrm{R}(\mathcal{Y})$ are distributed uniformly on $\mathbb{T}$. Since the original samples $\mathcal{Y}$ are distributed according to ϱ and we assume new test points are also samples according to ϱ, we are interested in the $L_2(\Omega, \varrho)$-error $\|\mathring{e}_f\|_{L_2(\Omega,\varrho)}$. Intuitively, if ϱ and $\mathring{\varrho}$ are equal enough, these two errors have the same behavior. This can be made more precise by

Theorem 6.1 *In the case where $\Omega = \mathbb{T}$, we have that*

$$\|\mathring{e}_f\|^2_{L_2(\Omega,\varrho)} \leqslant \|\mathring{e}_f\|^2_{L_2(\mathbb{T},\mathring{\varrho})} + \left(\int_{\mathbb{T}} |\mathring{e}_f(y)|^4 \,\mathrm{d}y\right)^{1/2} \|\mathring{\varrho}(y) - \varrho(y)\|_{L_2(\mathbb{T})}\,.$$

In the case $\Omega = \mathbb{R}$, there is a set $\mathcal{E} \subset \mathbb{R}$, such that $\int_{\mathcal{E}} \varrho(y)\mathrm{d}y \leqslant \varepsilon$ for some $\varepsilon > 0$. Therefore, we have

$$\|\mathring{e}_f\|^2_{L_2(\Omega,\varrho)} \leqslant \|\mathring{e}_f\|^2_{L_2(\mathbb{R},\mathring{\varrho})} + \left(\int_{\mathbb{R}\setminus\mathcal{E}} |\mathring{e}_f(y)|^4 \,\mathrm{d}y\right)^{1/2} \|\mathring{\varrho}(y) - \varrho(y)\|_{L_2(\mathbb{R})} + \varepsilon \sup_{y\in\mathcal{E}} |\mathring{e}_f|^2. \quad (6.3)$$

Proof The first inequality follows from triangle inequality and Cauchy-Schwarz inequality,

$$\begin{aligned}\|\mathring{e}_f\|^2_{L_2(\mathbb{T},\varrho)} &= \int_{-1/2}^{1/2} |\mathring{e}_f(y)|^2 \varrho(y)\,\mathrm{d}y = \int_{-1/2}^{1/2} |\mathring{e}_f(y)|^2 \mathring{\varrho}(y) \left(1 - \frac{\mathring{\varrho}(y) - \varrho(y)}{\mathring{\varrho}(y)}\right) \mathrm{d}y \\ &\leqslant \|\mathring{e}_f\|^2_{L_2(\mathbb{T},\mathring{\varrho})} + \int_{-1/2}^{1/2} |\mathring{e}_f(y)|^2 |\mathring{\varrho}(y) - \varrho(y)|\,\mathrm{d}y\end{aligned}$$

$$\begin{tikzcd}
\mathcal{Y} \subset \Omega \arrow[r, "(6.1)"] & \mathring{\varrho} \arrow[r, "(6.2)"] & \mathring{\mathrm{R}} \arrow[r, "\mathring{\mathcal{X}}=\mathring{\mathrm{R}}(\mathcal{Y})"] \arrow[drr] & \mathring{\mathcal{X}} \arrow[d] & \\
 & & \boldsymbol{f} \arrow[r] & S_n^{\mathring{\mathcal{X}}}(f \circ \mathring{\mathrm{R}}^{-1}) \arrow[r] & \left(S_n^{\mathring{\mathcal{X}}}(f \circ \mathring{\mathrm{R}}^{-1})\right) \circ \mathring{\mathrm{R}}
\end{tikzcd}$$

Fig. 14 Outline of our approximation procedure

$$\leqslant \|\mathring{e}_f\|^2_{L_2(\mathbb{T},\mathring{\varrho})} + \left(\int_{-1/2}^{1/2} |\mathring{e}_f(y)|^4 \,\mathrm{d}y\right)^{1/2} \|\mathring{\varrho}(y) - \varrho(y)\|_{L_2(\mathbb{T})} .$$

In the case where $\Omega = \mathbb{R}$, we can not use these calculations, since $\mathring{e}_f(y)$ can be nonzero on the whole real axis. But the splitting of $\mathbb{R}$ into the set $\mathcal{E}$ and the complement $\mathbb{R}\backslash\mathcal{E}$ and using the previous estimates gives

$$\begin{aligned}\|\mathring{e}_f\|^2_{L_2(\mathbb{R},\varrho)} &= \int_{\mathbb{R}\backslash\mathcal{E}} |\mathring{e}_f(y)|^2\varrho(y)\mathrm{d}y + \int_{\mathcal{E}} |\mathring{e}_f(y)|^2\varrho(y)\mathrm{d}y \\ &\leqslant \|\mathring{e}_f\|^2_{L_2(\mathbb{R}\backslash\mathcal{E},\mathring{\varrho})} + \left(\int_{\mathbb{R}\backslash\mathcal{E}} |\mathring{e}_f(y)|^4 \,\mathrm{d}y\right)^{1/2} \|\mathring{\varrho}(y) - \varrho(y)\|_{L_2(\mathbb{R}\backslash\mathcal{E})} + \varepsilon \sup_{y\in\mathcal{E}} |\mathring{e}_f|^2 \\ &\leqslant \|\mathring{e}_f\|^2_{L_2(\mathbb{R},\mathring{\varrho})} + \left(\int_{\mathbb{R}\backslash\mathcal{E}} |\mathring{e}_f(y)|^4 \,\mathrm{d}y\right)^{1/2} \|\mathring{\varrho}(y) - \varrho(y)\|_{L_2(\mathbb{R})} + \varepsilon \sup_{y\in\mathcal{E}} |\mathring{e}_f|^2.\end{aligned}$$

Therefore, it follows the assertion.

The introduction of the set $\mathcal{E}$ in the case where $\Omega = \mathbb{R}$ tackles the behavior, that given data $\mathcal{Y}$ is contained in a finite interval $[a, b]$ and $\varrho(y)$ is small outside this interval. In fact, we can not expect to approximate a function well where we have no information about the function.

The previous theorem shows that a good estimator for the bandwidth σ ensures that the mean *integrated squared error* (MISE)

$$\mathrm{MISE}(\mathring{\varrho}) = \mathbb{E}\left(\int_{\Omega} (\mathring{\varrho}(y) - \varrho(y))^2 \,\mathrm{d}y\right) = \mathbb{E}\left(\|\mathring{\varrho}(y) - \varrho(y)\|^2_{L_2(\Omega)}\right).$$

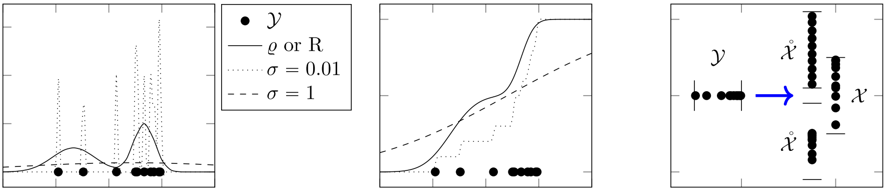
is small. This choice of the smoothing parameter has to be a good trade-off between over- and underfitting. Consider the two extremal cases, which do not work. This behavior is illustrated in Fig. 15.

- If we would choose a small smoothing parameter σ ($\sigma = 0.01$ in Fig. 15) or even as kernel function k the delta distribution, we would get an equi-spaced sample set $\mathring{\mathrm{R}}(\mathcal{Y})$. But this possesses on the other hand no smooth cumulative distribution function $\mathring{\mathrm{R}}$, and the distribution $\mathring{\varrho}$ is not a good approximation on ϱ, since $\mathrm{MISE}(\mathring{\varrho})$ would not decay.
- If we choose the parameter σ in the Kernel function k too big ($\sigma = 1$ in Fig. 15), the density $\mathring{\varrho}$ would not capture the behavior of the density ϱ and the transformed samples would not be uniformly distributed on $\mathbb{T}$.

6.1 Smoothing parameter selection

There are some very simple and easy-to-compute mathematical formulas for estimating the smoothing parameter σ. They are often called the *rules-of-thumb* (ROT). One possibility is

$$\sigma_{\mathrm{ROT}} = 1.06 \min\left\{\mathrm{std}(\mathcal{Y}), \frac{\mathrm{IQR}(\mathcal{Y})}{1.34}\right\} M^{-1/5}, \tag{6.4}$$



(a) The density function.

(b) The cumulative distribution function.

(c) The transformed samples.

Fig. 15 Illustration of the problem of over- and underfitting

where std is the *standard deviation* and IQR is the *inter-quartile range* (see [15, Section 4.2.1]). The assumption for that rule is that the unknown density belongs to the family of the normal distribution. In practice, we do not know, whether $\varrho(y)$ is a normal distribution. If it is, then σ_{ROT} gives the optimal smoothing parameter. If not, then σ_{ROT} will give a parameter not too far from the optimum, if the distribution of the samples $\mathcal{Y}$ is not too different from the normal distribution.

6.1.1 Data on the real axis

Another approach, which is more general and performs better than the ROT, is the *Direct Plug-In-Selector* (DPI) (see [15, Section 4.2.2]). To describe this approach, we have to introduce some notation. The second moment of the kernel k is defined by

$$\mu_2(k) = \int_{\mathbb{R}} y^2 k(y)\,\mathrm{d}y.$$

Since the MISE is set as the error criterion to be minimized, our aim is to find

$$\sigma_{\mathrm{MISE}} := \underset{\sigma>0}{\operatorname{argmin}}\, \mathrm{MISE}(\mathring{\varrho}).$$

The dominating part of MISE is denoted by AMISE, which stands for *Asymptotic MISE*,

$$\mathrm{AMISE}(\mathring{\varrho}) = \frac{1}{4}\mu_2^2(k)\,\|\varrho''\|^2_{L_2(\mathbb{R})}\,\sigma^4 + \frac{\|k\|^2_{L_2(\mathbb{R})}}{M\sigma},$$

The minimizer σ_{AMISE} is given by

$$\sigma_{\mathrm{AMISE}} = \left(\frac{\|k\|^2_{L_2(\mathbb{R})}}{\mu_2^2(k)\,\|\varrho''\|^2_{L_2(\mathbb{R})}\,M}\right)^{1/5} = \left(\frac{\|k\|^2_{L_2(\mathbb{R})}}{\mu_2^2(k)\Psi_4 M}\right)^{1/5}, \tag{6.5}$$

where

$$\Psi_4 = \int_{\mathbb{R}} \varrho^{(4)}(y)\varrho(y)\,\mathrm{d}y, \quad \text{or more generally } \Psi_r = \int_{\mathbb{R}} \varrho^{(r)}(y)\varrho(y)\,\mathrm{d}y,$$

where r is an even number. The naming convention of Ψ_r was introduced in [41, Section 3.5]. The critical step is to estimate Ψ_4 in (6.5), as this is the only unknown value. Assuming that ϱ is some normal distribution, would lead to (6.4). This is an example of a zero-stage plug-in selector, a terminology inspired by the fact that Ψ_4 was estimated by directly plugging in a parametric assumption. Another possibility is to estimate Ψ_4 non-parametrically and then to plug it into σ_{AMISE}. First note, that integration by parts gives

$$\|\varrho^{(r)}\|^2_{L^2(\mathbb{R})} = (-1)^r \int_{\mathbb{R}} \varrho^{(2r)}(y)\varrho(y)\,\mathrm{d}y.$$

Therefore, a possible way to estimate Ψ_r is

$$\overline{\Psi_r} = \frac{1}{M^2 g^{r+1}} \sum_{i=1}^{M}\sum_{j=1}^{M} k^{(r)}\left(\frac{y_i - y_j}{g}\right), \tag{6.6}$$

where g is the smoothing parameter of a kernel density estimation. Typically, two stages are considered to have a good trade-off between bias and variance. This is the method proposed by [38] and does the following steps.

i) Estimate Ψ_8 by $\overline{\Psi_8} = \left(\frac{105}{32\sqrt{\pi}(\overline{\mathrm{std}}(\varrho))^9}\right)$, where $\overline{\mathrm{std}}(\varrho)$ is an estimate for the standard derivation of ϱ, which can be $\mathrm{std}(\mathcal{X})$ or $\min\left\{\mathrm{std}(\mathcal{X}), \frac{\mathrm{IQR}(\mathcal{X})}{1.34}\right\}$.

ii) Estimate Ψ_6 using $\overline{\Psi_6}$ from (6.6), where $g_1 = \left(-\frac{2k^{(6)}(0)}{\mu_2(k)\overline{\Psi_8}M}\right)^{1/9}$.

iii) Estimate Ψ_4 using $\overline{\Psi_4}$ from (6.6), where $g_2 = \left(-\frac{2k^{(4)}(0)}{\mu_2(k)\overline{\Psi_6}M}\right)^{1/7}$.

iv) The selected smoothing parameter is $\sigma_{\mathrm{DPI}} := \left(\frac{\|k\|^2_{L_2(\mathbb{R})}}{\mu_2^2(k)\overline{\Psi_4}M}\right)^{1/5}$.

6.2 Non-periodic data

A general problem with kernel density estimation is that certain difficulties can arise at the boundaries and near them. In many practical situations, the values of a random variable are bounded. For example, the age of a person obviously can not be a negative number. On the other hand, the normal kernel has unlimited support. Even if a kernel with finite support is used, the estimated density can usually go beyond the permissible domain.

For that reason, we use on $\Omega = [0, 1]$ compactly supported kernels k, but we allow the approximated density $\mathring{\varrho}$ to be non-zero outside the interval $[0, 1]$. Especially, we receive a non-zero density in the interval $\mathring{\Omega} = [\omega_1, \omega_2]$ with $-|\operatorname{supp} k|\,\frac{\sigma}{2} \leqslant \omega_1 \leqslant 0$ and $1 \leqslant \omega_2 \leqslant 1 + |\operatorname{supp} k|\,\frac{\sigma}{2}$. This allows us to create a function $\mathring{f}$, which smoothly extends the function f to a function on whole $\mathring{\Omega}$, such that the transformed function $f \circ \mathring{\mathrm{R}}^{-1}$ becomes a periodic function. This idea of an extension of the function is similar to the studies in Sect. 5.1. The choice of the extension parameter η is now hidden in the choice of the smoothing parameter selection σ, which determines the interval $\mathring{\Omega}$. This is illustrated in Fig. 16. The kernel density estimation of ϱ can be seen as a periodization of the function f, such that we can use approximation operators for functions on $\mathbb{T}$. In contrast to the tent transformation (see [28]), which passes through the function forth and back, we have here no need to double the number of sample points.

To select the smoothing parameter σ, one simple possibility is to again use the estimator σ_{ROT} from (6.4). Analogously to Theorem 6.1, we give an estimate for the error decay, namely,

$$\|\mathring{e}_f\|^2_{L_2([0,1],\varrho)} = \int_0^1 |\mathring{e}_f(y)|^2 \frac{\varrho(y)}{\mathring{\varrho}(y)}\mathring{\varrho}(y)\mathrm{d}y \leqslant \left(\max_{y\in[0,1]} \frac{\varrho(y)}{\mathring{\varrho}(y)}\right) \|\mathring{e}_f\|^2_{L_2([0,1],\mathring{\varrho})} .$$

The extension $\mathring{f}$ is similar to the proposed method in Sect. 5.1. But here, naturally, the extension is on both sides of the interval because of the kernel density estimation. Instead of the factor $\frac{1}{1-\eta}$, we have now the the term $\max_{y\in[0,1]} \frac{\varrho(y)}{\mathring{\varrho}(y)}$.

6.3 Numerical experiments

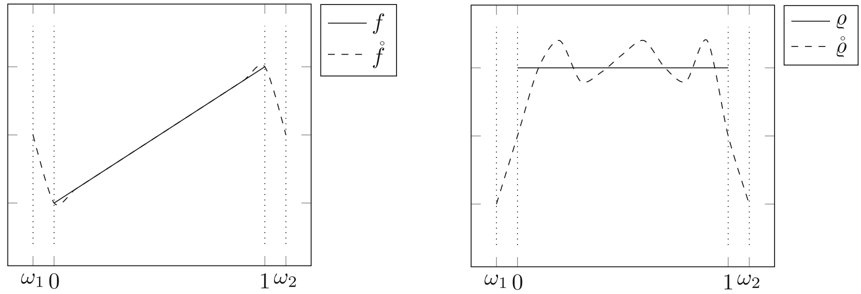
In this section, we endorse our theoretical findings by two numerical experiments on $\mathbb{R}$ and $[0, 1]$. We compare the approximation results of unknown density with the results if the density ϱ is known. In both cases, we use Chui-Wang wavelets of order $m = 3$.

The Gauss kernel on the real axis

For the case where $\Omega = \mathbb{R}$, we choose $k(y) = \varrho_N$ to be the standard normal distribution (4.5). The expressions used for estimating the smoothing parameter σ_{DPI} are

$$\|k\|^2_{L_2(\mathbb{R})} = \frac{1}{2\sqrt{\pi}}, \quad \mu_2(k) = 1.$$

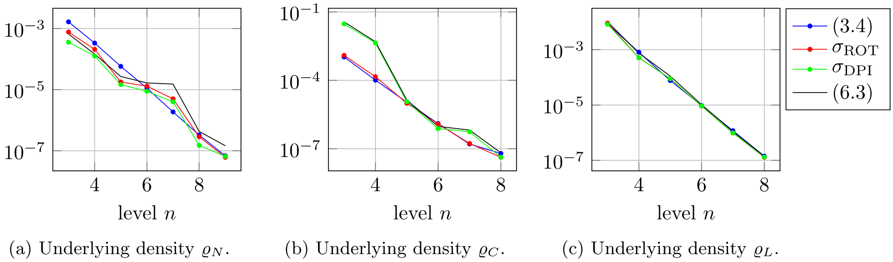
To study the performance of our algorithm, we use as a test function again the function in (5.13). We investigate the densities ϱ_N, (4.5); ϱ_C, (4.6); and ϱ_L, (4.7). Doing the same procedure as described in Sect. 5.2, we study the resulting RMSE (5.12). We used the two different proposed parameter selection methods and



(a) Function f and extended function $\mathring{f}$ on $\mathring{\Omega}$.

(b) The density ϱ and the estimated density $\mathring{\varrho}$.

Fig. 16 Periodization of f using instead of the real density ϱ the estimated density $\mathring{\varrho}$



(a) Underlying density ϱ_N. (b) Underlying density ϱ_C. (c) Underlying density ϱ_L.

Fig. 17 RMSE of the approximation on $\mathbb{R}$ of the test function (5.13) using kernel density estimation

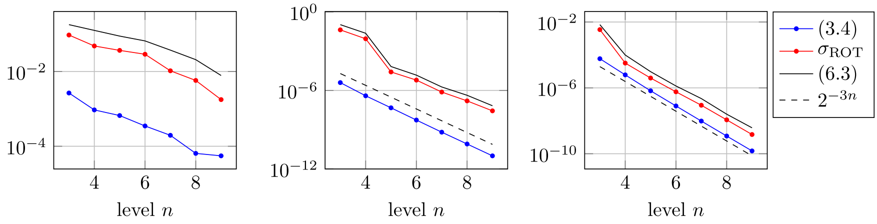
plotted the results in Fig. 17. It is reasonable to compare the results with the approximation error of the previous section, where we assume that the density ϱ is known (see (3.4) in Fig. 17). One can see that our approximation approach without knowing the density works well for all three examples, since for both investigated smoothing parameter selectors, we end up with nearly the same error, which we get with knowing the density. Furthermore, we had a look at the bound in Theorem 6.1. We choose

$$\mathcal{E} = (-\infty, -\max|\mathcal{Y}_{\text{test}}|] \cap [\max|\mathcal{Y}_{\text{test}}|, \infty),$$

since this is the interval where we do not expect data. Then, we calculated the right-hand side of (6.3) numerically for the choice σ_{DPI}. In Fig. 17, we see that this is indeed a good upper bound for the approximation error from choosing $\sigma = \sigma_{\text{DPI}}$.

The polynomial kernel on the cube

Let the density be the beta distribution $\varrho_{B,\alpha}$ from (4.20) with shape parameter $\alpha \in \{{}^1\!/\!_2, 1, 2\}$. Let us study the test function $f(y) = \mathrm{e}^y$. We use a B-Spline kernel $k(y) = B_3(y)$, see (B.1). In this case, the integral $K(y)$ can be calculated easily. The resulting RMSE are plotted in Fig. 18. We compare with the case where the density is known, (3.4). In the case where $\alpha = {}^1\!/\!_2$, the density $\varrho_{B,{}^1\!/\!_2}$ is large on the boundary (see Fig. 6a), which means that we have more points at the boundary. In this case, we receive the error rate 2^{-3n}, which is specified by the order of the wavelets.



(a) Beta distribution with parameter $\alpha = 2$. (b) Uniformly distributed samples on $[0, 1]$, i.e. $\alpha = 1$. (c) Beta distribution with parameter $\alpha = {}^1\!/\!_2$.

Fig. 18 RMSE of the approximation on the interval [0, 1] of the test function $f(y) = \mathrm{e}^y$ using kernel density estimation

This behavior occurs even in the case $\alpha = 1$, which means that the samples are uniformly distributed on $[0, 1]$. In the case where $\alpha = 2$, the benefit of our approximation approach is not as big as in the other cases, since the density ϱ tends to zero at the boundary and we do not have much samples at the boundary. This means that the function $\mathring{f}$ does not have support bigger than $[0, 1]$ and the smoothing effect of $\mathring{f} \circ \mathring{\mathrm{R}}^{-1}$ at the boundary does not apply. The approximation is slightly better in the case where the density is known, but nevertheless, this density does not inherit the decay rate from the other examples due to the lack of sample points at the boundary.

7 High-dimensional approximation

The main aim of this paper is the fast and effective approximation of high-dimensional functions. We study the setting where the variables $y_1, \dots, y_d$ are *independent*, which means that density $\varrho(y)$ is a product of the one-dimensional densities (3.7). Therefore, we transform every variable of the given samples separately after estimating one-dimensional densities. Additionally, we utilize the ANOVA decomposition from Sect. 3.1 to deal with the curse of dimensionality.

For the function $f \in L_2(\Omega, \varrho)$, we have the ANOVA decomposition (3.11). The number of ANOVA terms of a function is equal to 2^d and therefore grows exponentially in the dimension d. This reflects the curse of dimensionality in a certain way and poses a problem for the approximation of a function. In high-dimensional settings, the underlying function can very often be effectively represented as a sum of lower-order functions. In other words, the function can be expressed as a combination of component functions, where only $\nu \ll d$ variables out of the total d variables are active in each component [12, 25]. Recent methods such as ANOVAapprox [26, 31] (and the successful application to different datasets in [32]), SALSA [23], SRFE [17], and SHRIMP [44] use this approach. To this end, we introduce the notion of effective dimension (see [6]).

Definition 7.1 For $0 < \varepsilon \leqslant 1$, the *effective dimension* of f, in the *superposition sense*, is the smallest integer $\nu \leqslant d$, such that

$$\sum_{|\boldsymbol{u}| \leqslant \nu} \sigma_\varrho^2(f_{\boldsymbol{u}}) \geqslant \varepsilon \sigma_\varrho^2(f).$$

A function with low effective dimension allows a good approximation using only ANOVA terms up to order ν. To approximate $f_{\boldsymbol{u}}$, we have to use the transformation $\mathrm{R}_{\boldsymbol{u}}$, or if the density $\varrho_{\boldsymbol{u}}$ is unknown, we have to estimate the one-dimensional densities $\varrho_i(y_i)$ for $i \in \boldsymbol{u}$, and use a transformation $\mathring{\mathrm{R}}_{\boldsymbol{u}}(\boldsymbol{y}_{\boldsymbol{u}}) \approx \mathrm{R}_{\boldsymbol{u}}(\boldsymbol{y}_{\boldsymbol{u}})$ to transform the samples $\mathcal{Y}_{\boldsymbol{u}}$ to $\mathring{\mathcal{X}}_{\boldsymbol{u}} = \mathring{\mathrm{R}}_{\boldsymbol{u}}(\mathcal{Y}_{\boldsymbol{u}})$. Since we deal with independent input variables, we transform every variable separately, i.e.,

$$\mathring{\mathrm{R}}_{\boldsymbol{u}}(\boldsymbol{y}) = \left(\mathring{\mathrm{R}}_i(y_i)\right)_{i \in \boldsymbol{u}},$$

where we get $\mathring{\mathrm{R}}_i$ from one-dimensional transformations (6.2). In the truncated hyperbolic wavelet matrix $\boldsymbol{A}$, we insert only the indices $(\boldsymbol{j}, \boldsymbol{k})$ belonging to the low-dimensional terms, i.e.,

$$\boldsymbol{A} = \left(\psi^{\mathrm{per}}_{\boldsymbol{j},\boldsymbol{k}}(\boldsymbol{x}_{\boldsymbol{u}})\right)_{\boldsymbol{x}_{\boldsymbol{u}} \in \mathring{\mathcal{X}}_{\boldsymbol{u}}, (\boldsymbol{j},\boldsymbol{k}) \in I^{\boldsymbol{u}}_n}.$$

For chosen $\nu \ll d$, we do this for all $\boldsymbol{u}$ with $|\boldsymbol{u}| \leqslant \nu$, i.e., $U_\nu := \{\boldsymbol{u} \in [d] \mid |\boldsymbol{u}| \leqslant \nu\}$. This is summarized in Algorithm 1. The notation $\mathcal{Y}_{\{i\}}$ means analogously to the notation $\boldsymbol{y}_{\boldsymbol{u}}$ that we only consider the components y_i of the samples in $\mathcal{Y}$. Similar to [26, 31], we calculate the variances of the approximations of the ANOVA terms $f_{\boldsymbol{u}}$ and omit in a second approximation step the ones with low variance in order to increase the accuracy with a higher wavelet level n for the important ones. Hence, in a second approximation step, we use only the ANOVA terms $\boldsymbol{u} \in U \subset \mathcal{P}([d])$ and get the approximant $S^{\mathcal{Y},U}_n f$, (7.1), also for an arbitrary ANOVA index set U.

Algorithm 1 Transformed ANOVA hyperbolic wavelet regression.

```
Input:  $d$                                        dimension
        $\nu$                                      superposition dimension
        $\mathcal{Y} = (y_i)_{i=1}^M \in \Omega$   sampling nodes
        $\boldsymbol{f} = (f(y_i))_{i=1}^M$        function values at sampling nodes
1: Choose $n$ such that for $N = \sum_{|\boldsymbol{u}| \leqslant \nu} |I^{\boldsymbol{u}}_n|$ holds $M > N \log N$.
2: for $i = 1, \dots, d$ do
3:     if $\varrho_i(y_i)$ is known then
4:         Calculate the transformation $\mathrm{R}_i$ by (3.2).
5:         Transform the samples $\mathcal{Y}_i$ to $\mathcal{X}_i = \mathrm{R}_i(\mathcal{Y}_i)$.
6:         Do the steps 13-15 with $\mathrm{R}$ instead of $\mathring{\mathrm{R}}$.
7:     else
8:         Estimate $\mathring{\varrho}_i$ with (6.1).
9:         Calculate the transformation $\mathring{\mathrm{R}}_i$ by (6.2).
10:        Transform the samples $\mathcal{Y}_i$ to $\mathring{\mathcal{X}}_i = \mathring{\mathrm{R}}_i(\mathcal{Y}_i)$.
11:    end if
12: end for
13: Construct the sparse matrix

    $\boldsymbol{A} = [\boldsymbol{A}_{\boldsymbol{u}}]_{|\boldsymbol{u}| \leqslant \nu} \in \mathbb{C}^{M \times N}, \quad \boldsymbol{A}_{\boldsymbol{u}} = (\psi^{\mathrm{per}}_{\boldsymbol{j},\boldsymbol{k}}(\boldsymbol{x}_{\boldsymbol{u}}))_{\boldsymbol{x}_{\boldsymbol{u}} \in \mathring{\mathcal{X}}_{\boldsymbol{u}}, (\boldsymbol{j},\boldsymbol{k}) \in I^{\boldsymbol{u}}_n}.$

14: Solve the overdetermined linear system $\boldsymbol{A}\left(a_{\boldsymbol{j},\boldsymbol{k}}\right)_{\boldsymbol{j},\boldsymbol{k}} = \boldsymbol{f}$ via an LSQR-algorithm. This gives us the
    approximation

    $S^{\mathring{\mathcal{X}},\nu}_n (f \circ \mathring{\mathrm{R}})(\boldsymbol{x}) := \sum_{|\boldsymbol{u}| \leqslant \nu} \sum_{(\boldsymbol{j},\boldsymbol{k}) \in I^{\boldsymbol{u}}_n} a_{\boldsymbol{j},\boldsymbol{k}} \psi^{\mathrm{per}}_{\boldsymbol{j},\boldsymbol{k}}(\boldsymbol{x})$

15: Transform the approximation back to $\Omega$ using $\mathring{\mathrm{R}}^{-1}_{\boldsymbol{u}}$ for $|\boldsymbol{u}| \leqslant \nu$.
Output:  $\left(a_{\boldsymbol{j},\boldsymbol{k}}\right)_{\boldsymbol{j},\boldsymbol{k}} \in \mathbb{C}^N$ coefficients of the approximant

    $S^{\mathcal{Y},U}_n f(\boldsymbol{y}) := \sum_{\boldsymbol{u} \in U} \sum_{(\boldsymbol{j},\boldsymbol{k}) \in I^{\boldsymbol{u}}_n} a_{\boldsymbol{j},\boldsymbol{k}} \psi^{\mathrm{per}}_{\boldsymbol{j},\boldsymbol{k}}(\mathring{\mathrm{R}}^{-1}_{\boldsymbol{u}}(\boldsymbol{y}_{\boldsymbol{u}})),$    (7.1)

where $U = U_\nu$.
```

$$\begin{tikzcd}
\text{known } \varrho_i \arrow[r] & \text{transformation } \mathrm{R}_i \text{ from (3.2)} \arrow[r] & \mathcal{X}_i = \mathrm{R}(\mathcal{Y}_i) \\
\mathcal{Y}_i \arrow[u] \arrow[d] \arrow[urr] \arrow[drr] & & \\
\text{unknown } \varrho_i \arrow[r] & \text{transformation } \mathring{\mathrm{R}}_i \text{ from (6.2)} \arrow[r] & \mathring{\mathcal{X}}_i = \mathring{\mathrm{R}}(\mathcal{Y}_i)
\end{tikzcd}$$

Fig. 19 For every variable y_i of the sample set $\mathcal{Y}$, we have these possibilities

To summarize our algorithm, we have for every variable y_i two possibilities, which are summarized in Fig. 19. Algorithm 1 works well if the underlying density is a tensor density.

7.1 Approximating the global sensitivity indices

A direct calculation of sensitivity indices $S(\boldsymbol{u}, f)$ from (3.14) would require integral evaluations in (3.12), followed by numerous integral evaluations of sensitivity indices (3.13). For high-dimensional systems, such an approach is impractical and possibly prohibitive. Therefore, alternative routes must be charted to estimate the sensitivity indices both accurately and efficiently. Our approach is to approximate the function f by $S_n^{\mathcal{Y}} f$ and afterwards calculate the GSIs of the approximation. It was also shown in [16] that ANOVA terms inherit the smoothness of a function, i.e., if $f \in H^s_{\text{mix}}(\mathbb{T}^d)$, then $f_{\boldsymbol{u}} \in H^s_{\text{mix}}(\mathbb{T}^{|\boldsymbol{u}|})$ or even smoother. This fact was also shown in [31, Theorem 3.10] by using a Fourier-based approach. Following these lines, we immediately have this result for Besov spaces, i.e., if $f \in \boldsymbol{B}^s_{2,\infty}(\mathbb{T}^d)$ then $f_{\boldsymbol{u}} \in \boldsymbol{B}^s_{\text{mix}}(\mathbb{T}^{|\boldsymbol{u}|})$.

The intuition is that a good approximation of the function means also a good approximation of the ANOVA terms, and hence a good approximation of the variances. Calculating the variances of the ANOVA terms for functions on $\mathbb{T}^d$ is easy because of the connection (2.13). Therefore, we approximate the variances $\sigma^2_\varrho(f_{\boldsymbol{u}})$ by the following estimated variances:

$$\tilde{\sigma}^2_\varrho(f_{\boldsymbol{u}}) := \int_{\mathbb{T}^{|\boldsymbol{u}|}} |(S_n^{\mathcal{X}}(f \circ \mathrm{R}_{\boldsymbol{u}}^{-1}))_{\boldsymbol{u}}|^2 \,\mathrm{d}\boldsymbol{x}_{\boldsymbol{u}} = \int_{\Omega_{\boldsymbol{u}}} |(S_n^{\mathcal{X}}(f \circ \mathrm{R}_{\boldsymbol{u}}^{-1}))_{\boldsymbol{u}} \circ \mathrm{R}_{\boldsymbol{u}}|^2 \,\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{y}} \mathrm{R}_{\boldsymbol{u}}(\boldsymbol{y}_{\boldsymbol{u}}) \mathrm{d}\boldsymbol{y}_{\boldsymbol{u}},$$

$$\tilde{\sigma}^2_{\mathring{\varrho}}(f_{\boldsymbol{u}}) := \int_{\mathbb{T}^{|\boldsymbol{u}|}} |(S_n^{\mathring{\mathcal{X}}}(f \circ \mathring{\mathrm{R}}^{-1}))_{\boldsymbol{u}}|^2 \,\mathrm{d}\boldsymbol{x}_{\boldsymbol{u}} = \int_{\Omega_{\boldsymbol{u}}} |(S_n^{\mathring{\mathcal{X}}}(f \circ \mathring{\mathrm{R}}_{\boldsymbol{u}}^{-1}))_{\boldsymbol{u}} \circ \mathring{\mathrm{R}}_{\boldsymbol{u}}|^2 \,\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{y}} \mathring{\mathrm{R}}_{\boldsymbol{u}}(\boldsymbol{y}_{\boldsymbol{u}}) \mathrm{d}\boldsymbol{y}_{\boldsymbol{u}}$$

(see [26, Section 4] for computing details with hyperbolic wavelet regression). In the following, we will study the error between the estimated variances $\tilde{\sigma}^2_\varrho(f_{\boldsymbol{u}})$ and the variances $\sigma^2_\varrho(f_{\boldsymbol{u}})$. First, let us consider the case where the density is known.

Theorem 7.2 *Let* $\varnothing \neq \boldsymbol{u} \in U$ *and* $\boldsymbol{v} = \{i \in \boldsymbol{u} \mid \Omega_i = [0, 1]\}$ *and* $f_{\boldsymbol{u}} \in \boldsymbol{B}^s_{2,\infty}(\Omega_{\boldsymbol{u}}, \varrho_{\boldsymbol{u}})$. *Denote furthermore* $e_2 := \|f_{\boldsymbol{u}} - (S_n^{\mathcal{X}}(f \circ \mathrm{R}_{\boldsymbol{u}}^{-1})) \circ \mathrm{R}\|$ *and* $a = 1-(1-\eta)^{|\boldsymbol{v}|/2-1}$. *Then*

$$|\sigma^2_\varrho(f_{\boldsymbol{u}}) - \tilde{\sigma}^2_\varrho(f_{\boldsymbol{u}})| \leqslant \left(e_2 + a\,\|f_{\boldsymbol{u}}\|_{L_2(\Omega_{\boldsymbol{u}},\varrho_{\boldsymbol{u}})}\right) \left(2 + e_2 + a\,\|f_{\boldsymbol{u}}\|_{L_2(\Omega_{\boldsymbol{u}},\varrho_{\boldsymbol{u}})}\right) \|f_{\boldsymbol{u}}\|_{L_2(\Omega_{\boldsymbol{u}},\varrho_{\boldsymbol{u}})}\,.$$

Proof Let us denote in this proof $g = (S_n^{\mathcal{X}}(f \circ \mathrm{R}_{\boldsymbol{u}}^{-1})) \circ \mathrm{R}$. Then, we have

$$\tilde{\sigma}_{\varrho}^2(f_{\boldsymbol{u}}) = \int_{\Omega_{\boldsymbol{u}}} |g(\boldsymbol{y}_{\boldsymbol{u}})|^2 \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{y}} \mathrm{R}_{\boldsymbol{u}}(\boldsymbol{y}_{\boldsymbol{u}}) \mathrm{d}\boldsymbol{y}_{\boldsymbol{u}} = \int_{\Omega_{\boldsymbol{u}}} |g(\boldsymbol{y}_{\boldsymbol{u}})|^2 \, (1-\eta)^{-|\boldsymbol{v}|} \, \varrho_{\boldsymbol{u}}(\boldsymbol{y}_{\boldsymbol{u}}) \mathrm{d}\boldsymbol{y}_{\boldsymbol{u}}.$$

For simplicity, we denote in this proof $\tilde{g} = (1-\eta)^{-|\boldsymbol{v}|/2} g$. Then, we estimate the difference of the variances of the ANOVA terms given in (3.12) by the reverse triangle inequality and Cauchy-Schwarz inequality,

$$\begin{aligned} |\sigma_{\varrho}^2(f_{\boldsymbol{u}}) - \tilde{\sigma}_{\varrho}^2(f_{\boldsymbol{u}})| &= \left| \int_{\Omega_{\boldsymbol{u}}} \left(|f_{\boldsymbol{u}}(\boldsymbol{y}_{\boldsymbol{u}})|^2 - |\tilde{g}(\boldsymbol{y}_{\boldsymbol{u}})|^2 \right) \varrho_{\boldsymbol{u}}(\boldsymbol{y}_{\boldsymbol{u}}) \, \mathrm{d}\boldsymbol{y}_{\boldsymbol{u}} \right| \\ &= \left| \int_{\Omega_{\boldsymbol{u}}} \left(|f_{\boldsymbol{u}}(\boldsymbol{y}_{\boldsymbol{u}})| - |\tilde{g}(\boldsymbol{y}_{\boldsymbol{u}})| \right) \left(|f_{\boldsymbol{u}}(\boldsymbol{y}_{\boldsymbol{u}})| + |\tilde{g}(\boldsymbol{y}_{\boldsymbol{u}})| \right) \varrho_{\boldsymbol{u}}(\boldsymbol{y}_{\boldsymbol{u}}) \, \mathrm{d}\boldsymbol{y}_{\boldsymbol{u}} \right| \\ &\leqslant \|f_{\boldsymbol{u}} - \tilde{g}\|_{L_2(\Omega_{\boldsymbol{u}}, \varrho_{\boldsymbol{u}})} \, \|f_{\boldsymbol{u}} + \tilde{g}\|_{L_2(\Omega_{\boldsymbol{u}}, \varrho_{\boldsymbol{u}})} \\ &\leqslant \|f_{\boldsymbol{u}} - \tilde{g}\|_{L_2(\Omega_{\boldsymbol{u}}, \varrho_{\boldsymbol{u}})} \, \|f_{\boldsymbol{u}}\|_{L_2(\Omega_{\boldsymbol{u}}, \varrho_{\boldsymbol{u}})} \left(2 + \|f_{\boldsymbol{u}} - \tilde{g}\|_{L_2(\Omega_{\boldsymbol{u}}, \varrho_{\boldsymbol{u}})} \right). \end{aligned}$$

Only in the case where $\boldsymbol{v} \neq \varnothing$ we have the additional factor, which depends on η:

$$\|f_{\boldsymbol{u}} - \tilde{g}\|_{L_2(\Omega_{\boldsymbol{u}}, \varrho_{\boldsymbol{u}})} \leqslant \frac{1}{(1-\eta)^{|\boldsymbol{v}|/2}} \|f_{\boldsymbol{u}} - g\|_{L_2(\Omega_{\boldsymbol{u}}, \varrho_{\boldsymbol{u}})} + \left(1 - (1-\eta)^{|\boldsymbol{v}|/2 - 1}\right) \|f_{\boldsymbol{u}}\|_{L_2(\Omega_{\boldsymbol{u}}, \varrho_{\boldsymbol{u}})}$$

In the case of periodic functions, the second term is zero. Putting all inequalities together gives the desired result. □

The error between the function $f_{\boldsymbol{u}}$ and its approximation g can be estimated as follows. We are now concerned with a $|\boldsymbol{u}|$-dimensional function. Therefore, estimates (2.7) to (2.10) hold for $d = |\boldsymbol{u}|$ for the transformed function. The connection (3.10) or rather Theorem 5.1 transforms the results to $\Omega_{\boldsymbol{u}}$. Therefore, we have

$$\|f_{\boldsymbol{u}} - g\|_{L_2(\Omega_{\boldsymbol{u}}, \varrho_{\boldsymbol{u}})} \lesssim 2^{-2ns} n^{|\boldsymbol{u}|-1} \, \|f_{\boldsymbol{u}}\|_{\boldsymbol{B}_{2,\infty}^s},$$

with high probability. The logarithmic term in the approximation error appears only for ANOVA terms with $|\boldsymbol{u}| \geqslant 2$. In the case where the density ϱ is unknown, we get an additional term, which depends on the error between the estimated density $\mathring{\varrho}$ and the actual density ϱ, similar as in Theorem 6.1.
We introduce the subset

$$\mathcal{E} := \bigtimes_{i=1}^{d} \left(\Omega_i \backslash [\min \mathcal{X}_{\{i\}}, \max \mathcal{X}_{\{i\}}] \right),$$

for which the estimate in the next theorem follows.

Theorem 7.3 *Let $\varnothing \neq \boldsymbol{u} \in U$ and $f_{\boldsymbol{u}} \in \boldsymbol{B}^{s}_{2,\infty}(\Omega_{\boldsymbol{u}}, \varrho_{\boldsymbol{u}})$. Let furthermore $g := S_n^{\mathring{\mathcal{X}},U} f$ be an approximation received from our procedure in Algorithm 1 with unknown density. Then,*

$$|\sigma^2_{\varrho}(f_{\boldsymbol{u}}) - \tilde{\sigma}^2_{\varrho}(f_{\boldsymbol{u}})| \leqslant \underbrace{|\sigma^2_{\varrho}(f_{\boldsymbol{u}}) - \sigma^2_{\varrho}(g_{\boldsymbol{u}})|}_{A} + \underbrace{\left| \int_{\mathcal{E}} |g_{\boldsymbol{u}}(\boldsymbol{y})|^2 (\varrho(\boldsymbol{y}) - \mathring{\varrho}(\boldsymbol{y}))\,\mathrm{d}\boldsymbol{y} \right|}_{B}$$

$$+ \underbrace{\left| \int_{\Omega_{\boldsymbol{u}} \setminus \mathcal{E}} |g_{\boldsymbol{u}}(\boldsymbol{y})|^2 \varrho(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} \right| \sup_{\boldsymbol{y} \in \Omega_{\boldsymbol{u}} \setminus \mathcal{E}} \left(1 - \frac{\mathring{\varrho}(\boldsymbol{y})}{\varrho(\boldsymbol{y})} \right)}_{C}.$$

Let us briefly explain the appearing terms. Term A is the error from Theorem 7.2, term C depends on the quality of the approximation $\mathring{\varrho} \approx \varrho$, and term B describes the variance of $g_{\boldsymbol{u}}$ in the part where we have no samples, i.e., where we extend the original function $f_{\boldsymbol{u}}$. Of course, if the original function is non-periodic, we use an extension and study the variances, so we have to accept the additional term.

Proof We do the following splitting:

$$|\sigma^2_{\varrho}(f_{\boldsymbol{u}}) - \sigma^2_{\mathring{\varrho}}(g_{\boldsymbol{u}})| \leqslant |\sigma^2_{\varrho}(f_{\boldsymbol{u}}) - \sigma^2_{\varrho}(g_{\boldsymbol{u}})| + |\sigma^2_{\varrho}(g_{\boldsymbol{u}}) - \sigma^2_{\mathring{\varrho}}(g_{\boldsymbol{u}})|.$$

Then, by splitting the domain of the second integral, the assertion follows. □

The quintessence of this subsection is that the approximation of the GSI of the function f by the GSI of the approximant is a reasonable approach to get insides about the variances of the ANOVA terms. In a second approximation step, we reduce the index set to the ANOVA indices in the set

$$U = \{\boldsymbol{u} \in U_\nu \mid S(\boldsymbol{u}, S_n^{\mathcal{Y}} f) > \varepsilon\},$$

for some threshold $\varepsilon > 0$. This allows on the other hand to increase the maximal wavelet level for the important ANOVA terms and therefore decrease the approximation error, while ensuring logarithmic oversampling.

7.2 A synthetic numerical example

As a conclusion of this paper, we want to apply Algorithm 1 to a high-dimensional test function. For that reason, let

$$f\colon \mathbb{R}^5 \times [0,1]^3 \to \mathbb{R}, \quad f(\boldsymbol{y}) = \tfrac{1}{5} y_1^2 + \tfrac{1}{2}\cos(2\pi y_3) + \mathrm{e}^{-y_4^2} + y_5^{1/2} + 30\,(y_6^3\,(1 - y_6^2)) + \tfrac{1}{2}|4\,y_7 - 2| + 5\mathrm{e}^{-y_1^2 - y_5^2} \tag{7.2}$$

be an 8-dimensional function where $y_5 > 0$. We assume the given data $\mathcal{Y}$ to be sampled from the distribution

$$\varrho\colon \mathbb{R}^5 \times [0,1]^3 \to \mathbb{R}_+, \quad \varrho(\boldsymbol{y}) = \prod_{i=1}^{8} \varrho_i(y_i),$$

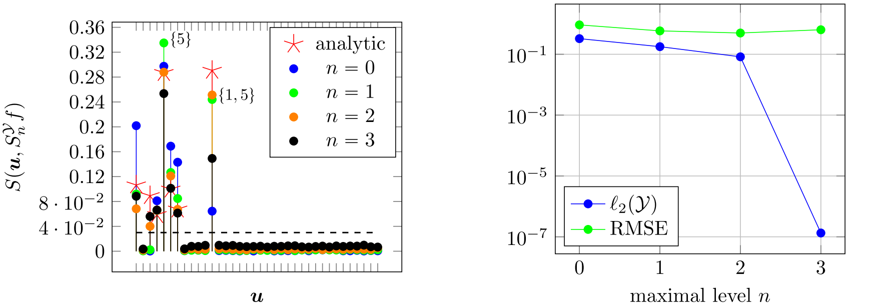
where we use different distributions, already studied in this paper,

$$\varrho_1(y_1) = \varrho_N(y_1), \quad \varrho_2(y_2) = \varrho_L(y_2), \quad \varrho_3(y_3) = \varrho_C(y_3), \quad \varrho_5(y_5) = \frac{1}{2}\mathrm{e}^{-\frac{y_5}{2}}, \quad \varrho_6(y_6) = \varrho_8(y_8) = 1,$$

$$\varrho_4(y_4) = \frac{1}{\sqrt{11.52\,\pi}}\mathrm{e}^{-\frac{(y_4+2)^2}{2.88}} + \frac{1}{\sqrt{50\,\pi}}\mathrm{e}^{-\frac{(y_4-3)^2}{12.5}}, \quad \varrho_7(y_7) = \frac{1}{\pi}\, y_7{}^{-1/2}(1-y_7)^{-1/2}.$$

We draw $M = 1000$ samples and use the corresponding function values $\boldsymbol{f} = f(\mathcal{Y})$. These samples projected to the directions y_1 and y_2 are plotted in the introduction in Fig. 1 together with the transformed samples $\mathring{\mathrm{R}}(\mathcal{Y})$, also projected to the directions y_1 and y_2. We use as superposition dimension $\nu = 2$, which is a suitable guess if we have a look at the function equation, which suggests only one ANOVA term of order 2. With the choice $\nu = 3$, we would conclude in a first step that we do not need the three-dimensional terms. Furthermore, we use Chui-Wang wavelets of order $m = 2$. We consider the setting where we do not know the underlying density, so we use for the variables $y_1, \dots, y_5$ the kernel density estimation for data on $\mathbb{R}$ from Sect. 6.1.1 using the Gaussian kernel introduced there and for the remaining variables y_6, y_7, y_8 the kernel density estimation for data on $[0, 1]$ from Sect. 6.2 using the B-spline kernel introduced there. For different wavelet levels n, we plot in Fig. 20a the approximated GSI's $S(\boldsymbol{u}, S_n^{\mathcal{Y}} f)$, i.e., the 8 GSI's of order 1 and then the 28 GSI's of order 2. Since we know the function explicitly, we compare this to the analytically calculated GSI's $S(\boldsymbol{u}, f)$. One can see that we could indeed figure out even with a low wavelet level $n = 2$ the ANOVA terms with high variances. So we filter out the unnecessary variables y_2 and y_8 and all two-dimensional terms except the term with $\boldsymbol{u} = \{1, 5\}$. Furthermore, we plotted in Fig. 20b the error $\|f - S_n^{\mathcal{Y}} f\|_{\ell_2(\mathcal{Y})} = \left(\sum_{\boldsymbol{y}\in\mathcal{Y}} |f(\boldsymbol{y}) - S_n^{\mathcal{Y}} f(\boldsymbol{y})|^2\right)^{1/2}$ and the RMSE (5.12) for a test set $\mathcal{Y}_{\text{test}}$ sampled according to ϱ with $|\mathcal{Y}_{\text{test}}| = 3M$. The low $\ell_2(\mathcal{Y})$-error indicates that $n = 3$ is already overfitting, i.e., using too many parameters for the 1000 samples.

In a second approximation step, we omit the ANOVA terms with low variance and use only the ANOVA indices $U = \{\varnothing, \{1\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7\}, \{1, 5\}\}$ for the approximation. This procedure is similar to the proposed method suggested in [26,



(a) The approximated GSIs $S(\boldsymbol{u}, S_n^{\mathcal{Y}} f)$ for $|\boldsymbol{u}| \leqslant 2$ and different level n compared to the GSIs $S(\boldsymbol{u}, f)$.

(b) Approximation errors, the ℓ_2-norm on the given samples and the RMSE for different wavelet levels n.

Fig. 20 Approximation of the function f from (7.2) from $M = 1000$ samples

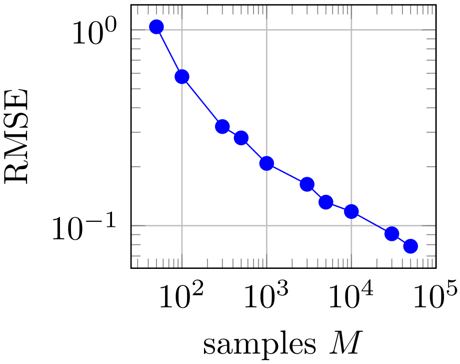


Fig. 21 Approximation of the function f from (7.2) for different numbers of samples M

32]. It is also reasonable to choose different maximal levels n for the one- and the two-dimensional terms, since for different dimensions, these index sets have different sizes. For the choice $n = 5$ for the one-dimensional terms and the choice $n = 3$ for the two-dimensional term, we are able to reduce the RMSE on the test set additionally from 0.4997 to 0.2248. Our procedure reduces the RMSE significantly; hence, we are able to approximate an eight-dimensional function using only 1000 samples very well. Using only a min-max transformation of the data, it is not possible to detect the non-zero ANOVA terms.

In a second experiment, we only use the ANOVA indices in U for the approximation and do our procedure for different sample sizes M, with all other parameters kept the same. The results are plotted in Fig. 21. We used the maximal wavelet level n for which the RMSE is minimal for the given fixed data set.

7.3 Real-world data

The proposed Algorithm 1 was used in [42] to estimate the vertical ground reaction force from time series of plantar pressure from instrumented insoles. The study included data from 16 persons moving at different speeds on a two-belt treadmill equipped with force plates. In total, about $M \approx 1.2 \cdot 10^6$ data points were used and the data was modelled as an 8-dimensional function with ANOVA terms up to order $\nu = 2$. The approach successfully reached relative RMSEs of up to 10.6 %, which is comparable to other studies in the literature with the advantage of being interpretable and having much more data available in the study used in [42].

In the following, we compare the performance of models obtained by Algorithm 1 with other state-of-the-art algorithms when applied to seven real-world datasets from the the UCI repository (http://archive.ics.uci.edu/ml). For comparison, we test against random forest regression (RF) and Gaussian processes (GP), both implemented in ScikitLearn.jl, which implements the popular scikit-learn interface and algorithms in Julia. Furthermore, we compare with sparse additive models [23, 36, 44] and follow the experimental setup in [23], where the training data is normalized so that the input and output values have zero mean and unit variance along each dimension. Each dataset is divided in half to form the training and testing sets, and we use exactly the same splitting as in [23] for the datasets used there and do the same procedure for the

Table 2 Overview of seven datasets: dimension d and number of datapoints in training and testing set

Dataset	d	M_{train}	M_{test}	Basis	N
Propulsion	15	200	200	chui-2	271
Galaxy	20	2000	2000	per	983
Airfoil	41	750	750	chui-2	339
Forestfire	10	211	167	per	100
Boston Housing	12	256	250	chui-2	166
Protein	9	22, 685	22, 685	per	2082
Elevators	17	8300	8399	chui-4	2332

The experimental setup and datasets for each test follow from [23]. We give also the used basis functions on the torus, the total number N of trained coefficients in the final model of Algorithm 1

datasets with more samples. Note that they test only one single random splitting. In our experiments, we use a cross-validation of the original training dataset as training data and 20% as validation data to select the best parameters ν, n, λ, ε. An overview of the datasets is presented in Table 2. Furthermore, we give details of our trained model: The used basis functions on the torus (chui-m are Chui-Wang-wavelets of order m and per means trigonometrical polynomials) as well as the total number N of trained coefficients in the final model. For the transformation R, we use the DPI, described in Sect. 6.1.1 applied to the Gaussian kernel.

The approximation results and comparisons are shown in Table 3. The results of SALSA are obtained from [23], the results from HARFE and SRFE are obtained from [36], and we included the results of the SHRIMP algorithm [44]. Since the results for the random algorithms depend on the draw of the random features, in contrast to the given results for SHRIMP, HARFE, and SALSA, we did the approximation validation 50 times and present the mean in Table 3. Note that our parameter ν coincides with the parameter q in the random feature literature. Furthermore, for the datasets with too many samples, i.e., Protein and Elevators, the random feature algorithms are not able to calculate an approximation, since the involved random matrices are getting too big.

Table 3 MSE on real datasets using various approximation techniques

	Propulsion	Galaxy	Airfoil	Forestfire	Housing	Protein	Elevators
Alg. 1 (ν)	**0.0001126** (2)	0.00344 (1)	**0.1530** (2)	0.3460 (1)	0.3779 (1)	**0.4095** (3)	**0.2488** (1)
RF	0.005928	0.1092	0.6358	0.3372	0.3339	0.4225	0.2908
GP	0.009031	0.02765	1.0091	0.4729	0.4231	0.4414	0.4324
HARFE (ν)	0.000140 (2)	**0.000110** (2)	0.5350 (2)	**0.3122** (2)	**0.2994** (2)	-	-
SHRIMP (ν)	0.000147 (1)	0.000190 (2)	0.3616 (2)	0.3501 (1)	0.4551 (7)	-	-
SALSA	0.00918	0.000135	0.5470	0.3635	0.3607	-	-
SRFE	0.0154	0.0012	0.5702	0.4067	0.6395	-	-

Details to the corresponding algorithms can be found in HARFE [36], SHRIMP [44], SALSA [23], and SRFE [17] or are available via ScikitLearn.jl. The best results for every dataset are highlighted

We want to highlight two cases: the cases where the dataset has many samples M compared to the dimension d, and second, the dataset has not much data samples available. In the first case (datasets Protein and Elevators), Algorithm 1 performs better than the other machine learning algorithms. Since the random feature models set up a matrix as large as the number of unknowns, they can not handle such big datasets. Our Algorithm 1 even provides similar or smaller approximation errors compared to the random features models in the case of a low sample size, with the advantage of being interpretable, that means it is easy to calculate the GSIs of the involved ANOVA terms for the final approximation model. In the application, the user can use this to derive conclusions. Furthermore, we confirm the thesis, that real-world data can be described by functions with low effective dimension. It should also be noted that in the dataset Airfoil, additional 36 dimensions with random noise are added and Algorithm 1 easily finds the non-importance of these dimensions and reduces the data to the important 5 dimensions.

8 Conclusion and outlook

In this paper, we introduced a new method for function approximation from given fixed samples from an arbitrary density. This method combines previous work on least squares approximation on the torus $\mathbb{T}^d$ and the truncation of the ANOVA decomposition with a variable transformation and a kernel density estimation. We are able to transfer the error decay rates and the fast algorithms from the torus to the domain Ω. The new extension method, which benefits from the Chui-Wang wavelets, even allows the approximation of non-periodic functions. As shown in our numerical experiments, this procedure is beneficial in function approximation. The code is available in the Julia package ANOVAapprox on GitHub (see https://github.com/NFFT/ANOVAapprox). We assume for our algorithm that the input variables are independent, which is not necessarily the case in applications. In future work, we want to adapt our algorithm also to the setting of dependent input variables.

Appendix 1. Besov-Nikolskij-Sobolev spaces of mixed smoothness on the *d*-torus

Here, we summarize some relevant results from [10, Chapt. 3]. In particular, we give the standard definition of the used function spaces. Let us first define Besov-Nikolskij spaces of mixed smoothness. We will use the classical definition via mixed moduli of smoothness. Therefore, first, recall the basic concepts. For univariate functions $f : \mathbb{T} \to \mathbb{C}$, the m-th difference operator Δ_h^m is defined by

$$\Delta_h^m(f, x) := \sum_{j=0}^{m} (-1)^{m-j} \binom{m}{j} f(x + jh), \quad x \in \mathbb{T}, h \in [0, 1] .$$

Let $\boldsymbol{u}$ be any subset of $\{1, ..., d\}$. For multivariate functions $f : \mathbb{T}^d \to \mathbb{C}$ and $\boldsymbol{h} \in [0,1]^d$, the mixed $(m, \boldsymbol{u})$-th difference operator $\Delta_{\boldsymbol{h}}^{m,\boldsymbol{u}}$ is defined by

$$\Delta_{\boldsymbol{h}}^{m,\boldsymbol{u}} := \prod_{i\in\boldsymbol{u}} \Delta_{h_i,i}^{m} \quad \text{and} \quad \Delta_{\boldsymbol{h}}^{m,\varnothing} = \mathrm{Id},$$

where $\mathrm{Id}\, f = f$ and $\Delta_{h_i,i}^{m}$ is the univariate operator applied to the i-th coordinate of f with the other variables kept fixed.

Definition A.1 Let $s > 0$ and $1 \le p \le \infty$. Fixing an integer $m > s$, we define the space $\boldsymbol{B}_{p,\infty}^{s}(\mathbb{T}^d)$ as the set of all $f \in L_p(\mathbb{T}^d)$ such that for any $\boldsymbol{u} \subset \{1, ..., d\}$

$$\left\|\Delta_{\boldsymbol{h}}^{m,\boldsymbol{u}}(f,\cdot)\right\|_{L_p(\mathbb{T}^d)} \le C \prod_{i\in\boldsymbol{u}} |h_i|^s$$

for some positive constant C and introduce the norm in this space

$$\| f \|_{\boldsymbol{B}_{p,\infty}^{s}} := \sum_{\boldsymbol{u}\subseteq\{1,...,d\}} | f |_{\boldsymbol{B}_{p,\infty}^{s}(\boldsymbol{u})},$$

where

$$| f |_{\boldsymbol{B}_{p,\infty}^{s}(\boldsymbol{u})} := \sup_{0<|h_i|\le 1,\ i\in\boldsymbol{u}} \left(\prod_{i\in\boldsymbol{u}} |h_i|^{-s}\right) \| \Delta_{\boldsymbol{h}}^{m,\boldsymbol{u}}(f,\cdot) \|_{L_p(\mathbb{T}^d)} .$$

We define functions in a Sobolev space with dominating mixed derivatives,

$$H_{\mathrm{mix}}^{m}(\mathbb{T}^d) := \left\{ f : \mathbb{T}^d \to \mathbb{C} \mid \|f\|_{H_{\mathrm{mix}}^{m}(\mathbb{T}^d)} < \infty \right\},$$

where the norm is defined by

$$\|f\|_{H_{\mathrm{mix}}^{m}(\mathbb{T}^d)} = \sum_{0\leqslant\|\boldsymbol{k}\|_\infty\leqslant m} \left\|\mathrm{D}^{\boldsymbol{k}} f\right\|_{L_2(\mathbb{T}^d)}, \tag{A.1}$$

with the partial derivatives $\mathrm{D}^{\boldsymbol{k}} f = \frac{\partial^{k_1+...+k_d}}{\partial x_1^{k_1} \cdots \partial x_d^{k_d}}$. It clearly holds for $d = 1$

$$H_{\mathrm{mix}}^{m}(\mathbb{T}) = H_2^m(\mathbb{T}) =: H^m(\mathbb{T}) .$$

The case $p = 2$ and $\Omega = \mathbb{T}^d$ allows for a straightforward extension to fractional smoothness parameters.

Definition A.2 Let $s > 0$. Then, we define

$$H_{\mathrm{mix}}^{s}(\mathbb{T}^d) := \left\{ f : \mathbb{T}^d \to \mathbb{C} \mid \|f\|_{H_{\mathrm{mix}}^{s}(\mathbb{T}^d)} < \infty \right\},$$

where the norm is defined by

$$\|f\|^2_{H^s_{\mathrm{mix}}(\mathbb{T}^d)} = \sum_{\boldsymbol{k}\in\mathbb{Z}^d} |c_{\boldsymbol{k}}(f)|^2 \prod_{i=1}^{d} (1+|k_i|^2)^s. \tag{A.2}$$

This norm is equivalent to the norm in (A.1) for $s \in \mathbb{N}$ (see [24]). We will consider the case where $s > \frac{1}{2}$, since in this case we have that $H^s_{\mathrm{mix}}(\mathbb{T}^d) \hookrightarrow C(\mathbb{T}^d)$, which is necessary to sample the function. There is a further useful equivalent norm which is based on a decomposition of f in dyadic blocks. The dyadic blocks (4.3) and the decomposition (4.16) immediately give

$$\|f\|^2_{H^s_{\mathrm{mix}}} \asymp \sum_{\boldsymbol{j}\in\mathbb{N}_0^d} 2^{2|\boldsymbol{j}|_1 s} \|f_{\boldsymbol{j}}\|^2_{L_2(\mathbb{T}^d)}.$$

Interestingly, there is also a Fourier analytic characterization of the above-defined Besov-Nikolskij spaces $\boldsymbol{B}^s_{p,\infty}(\mathbb{T}^d)$ which even works for $1 < p < \infty$. Instead of taking the ℓ_2-norm of the weighted sequence $(2^{|\boldsymbol{j}|_1 s}\|f_{\boldsymbol{j}}\|_{L_p(\mathbb{T}^d)})_{\boldsymbol{j}\in\mathbb{N}_0^d}$, we take the ℓ_∞-norm,

$$\|f\|_{\boldsymbol{B}^s_{p,\infty}} \asymp \sup_{\boldsymbol{j}\in\mathbb{N}_0^d} 2^{|\boldsymbol{j}|_1 s} \|f_{\boldsymbol{j}}\|_{L_p(\mathbb{T}^d)}\,. \tag{A.3}$$

Appendix 2. Cardinal B-splines and Chui-Wang wavelets

We mostly work in this paper with spline wavelets, which have useful properties for our purposes. Therefore, we introduce here the cardinal B-splines and the corresponding Chui-Wang wavelets. The cardinal B-spline $B_m : \mathbb{R} \to \mathbb{R}$ of order m is a piecewise polynomial function recursively defined by

$$B_1(x) = \begin{cases} 1, & -\frac{1}{2} < x < \frac{1}{2} \\ 0, & \text{otherwise,} \end{cases} \quad \text{and} \quad B_m(x) = \int_{x-1/2}^{x+1/2} B_{m-1}(t)\,\mathrm{d}t. \tag{B.1}$$

The function $B_m(x)$ is a piecewise polynomial function of order $m-1$. Furthermore, the support of $B_m(x)$ is $(-\frac{m}{2}, \frac{m}{2})$, and they are normalized by $\int_{-m/2}^{m/2} B_m(x)\mathrm{d}x = 1$.

Definition B.1 If we use the cardinal B-spline of order m as scaling function $\varphi = B_m$, the corresponding wavelet functions are the Chui-Wang wavelets [7], which are given by

$$\psi(x) = \sum_n q_n B_m(2x - n - \tfrac{m}{2}),$$

where

$$q_n = \frac{(-1)^n}{2^{m-1}} \sum_{k=0}^{m} \binom{m}{k} B_{2m}(n+1-k-m).$$

As in [26], we introduce the function

$$\Psi_m(x) := \int_{-\infty}^{x} \frac{\psi(t)(x-t)^{m-1}}{(m-1)!} \,\mathrm{d}t, \tag{B.2}$$

which is supported on $[0, 2m-1]$ and fulfills $\mathrm{D}^m \Psi_m(x) = \psi(x)$ and is bounded, i.e., there exists a constant C_m, such that

$$|\Psi_m(x)| \leqslant C_m \text{ for all } x \in \mathbb{R}. \tag{B.3}$$

Furthermore, this function has a nice property, which we use in Lemma 5.5.

Lemma B.2 *For $k \in \mathbb{Z}$, we have for the function* (B.2) *that*

$$\Psi_m(k) = 0.$$

Proof The result is a consequence of results from [7, Chapter 6]. The wavelet function ψ can also be written as

$$\psi(x) = L_{2m}^{(m)}(2x-1),$$

where $L_m(x)$ is the fundamental cardinal spline, which is a piecewise polynomial of degree m with the interpolation property $L_m(j) = \delta_{j,0}$ for $j \in \mathbb{Z}$. Then, we have $\Psi_m(x) = \frac{1}{2^m} L_{2m}(2x-1)$, which is zero for $x \in \mathbb{Z}$. □

Acknowledgements The authors thank the referees for useful comments about this work. The authors would like to thank Tino Ullrich for several fruitful discussions.

Funding Open Access funding enabled and organized by Projekt DEAL. The authors acknowledge funding by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project-ID 416228727 - SFB 1410. Furthermore, the second author acknowledges the support by the BMBF grant 01|S20053A (project SAℓE).

Data Availability The code is available in the Julia package ANOVAapprox on GitHub, see https://github.com/NFFT/ANOVAapprox. The data sets generated during and/or analysed during the current study are available from the corresponding author on request.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Adcock, B., Brugiapaglia, S., Webster, C.G.: Sparse polynomial approximation of high-dimensional functions. Computational science & engineering. SIAM, Philadelphia, Pennsylvania (2022)
2. Adcock, B., Huybrechs, D.: Frames and numerical approximation. SIAM Review **61**(3), 443–473 (2019)
3. Adcock, B., Huybrechs, D.: Approximating smooth, multivariate functions on irregular domains. Forum Math. Sigma **8**, E26 (2020)
4. Bell, E.T.: Exponential polynomials. Ann. Math **35**(2), 258–277 (1934)
5. Boyd, J.P.: Six strategies for defeating the Runge phenomenon in Gaussian radial basis functions on a finite interval. Comput. Math. Appl. **60**(12), 3108–3122 (2010)
6. Caflisch, R., Morokoff, W., Owen, A.: Valuation of mortgage-backed securities using Brownian bridges to reduce effective dimension. J. Comput. Finance **1**(1), 27–46 (1997)
7. Chui, C.K.: An introduction to wavelets. Academic Press, Boston (1992)
8. Cohen, A., Davenport, M.A., Leviatan, D.: On the stability and accuracy of least-squares approximations. Found. Comput. Math. **13**, 819–834 (2013)
9. Cohen, A., Migliorati, G.: Optimal weighted least-squares methods. SMAI J. Comput. Math. **3**, 181–203 (2017)
10. Dũng, D., Temlyakov, V.N., Ullrich, T.: Hyperbolic cross approximation. Advanced Courses in Mathematics - CRM Barcelona. Birkhäuser, Cham (2018)
11. Dahmen, W., Kunoth, A., Urban, K.: Biorthogonal spline wavelets on the interval-stability and moment conditions. Appl. Comput. Harmon. Anal. **6**(2), 132–196 (1999)
12. DeVore, R., Petrova, G., Wojtaszczyk, P.: Approximation of functions of few variables in high dimensions. Constr. Approx. **33**(1), 125–143 (2010)
13. Dolbeault, M., Cohen, A.: Optimal pointwise sampling for l^2 approximation. J. Complex. **68**, 101602 (2022)
14. Gilbert, A.D., Kuo, F.Y., Sloan, I.H.: Equivalence between Sobolev spaces of first-order dominating mixed smoothness and unanchored ANOVA spaces on $\mathbb{R}^d$. Math. Comp. **91**, 1837–1869 (2022)
15. Gramacki, A.: Nonparametric kernel density estimation and its computational aspects, vol. 37. Springer International (2018)
16. Griebel, M., Kuo, F.Y., Sloan, I.H.: The smoothing effect of the ANOVA decomposition. J. Complex. **26**(5), 523–551 (2010)
17. Hashemi, A., Schaeffer, H., Shi, R., Topcu, U., Tran, G., Ward, R.: Generalization bounds for sparse random feature expansions. Appl. Comput. Harmon. Anal. **62**, 310–330 (2023)
18. Holtz, M.: Sparse grid quadrature in high dimensions with applications in finance and insurance. Lecture Notes in Computational Science and Engineering, vol. 77. Springer-Verlag, Berlin (2011)
19. Hooker, G.: Generalized functional ANOVA diagnostics for high-dimensional functions of dependent variables. J. Comput. Graph. Stat. **16**(3), 709–732 (2007)
20. Huybrechs, D.: On the Fourier extension of nonperiodic functions. SIAM J. Numer. Anal. **47**(6), 4326–4355 (2010)
21. Jia, R.-Q.: Spline wavelets on the interval with homogeneous boundary conditions. Adv. Comput. Math. **30**, 177–200 (2009)
22. Kämmerer, L., Ullrich, T., Volkmer, T.: Worst case recovery guarantees for least squares approximation using random samples. Constr. Approx. **54**, 295–352 (2021)
23. Kandasamy, K., Yu, Y.: Additive approximations in high dimensional nonparametric regression via the salsa. Int Conf Mach Learn 69–78 (2016). PMLR
24. Kühn, T., Sickel, W., Ullrich, T.: Approximation numbers of Sobolev embeddings – sharp constants and tractability. J. Complex. **30**, 95–116 (2014)
25. Kuo, F.Y., Sloan, I.H., Wasilkowski, G.W., Woźniakowski, H.: On decompositions of multivariate functions. Math. Comp. **79**(270), 953–966 (2010)
26. Lippert, L., Potts, D., Ullrich, T.: Fast hyperbolic wavelet regression meets ANOVA. Numer. Math. **154**, 155–207 (2023)
27. Liu, R., Owen, A.B.: Estimating mean dimensionality of analysis of variance decompositions. J. Amer. Statist. Assoc. **101**(474), 712–721 (2006)
28. Nasdala, R.: Efficient multivariate approximation with transformed rank-1 lattices. Dissertation, Fakultät für Mathematik, Technische Universität Chemnitz (2021)

29. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems Volume I: Linear Information. Eur. Math. Society, EMS Tracts in Mathematics **6** (2008)
30. Nuyens, D., Suzuki, Y.: Scaled lattice rules for integration on $\mathbb{R}^d$ achieving higher-order convergence with error analysis in terms of orthogonal projections onto periodic spaces. Math. Comp. **92**, 307–347 (2023)
31. Potts, D., Schmischke, M.: Approximation of high-dimensional periodic functions with Fourier-based methods. SIAM J. Numer. Anal. **59**(5), 2393–2429 (2021)
32. Potts, D., Schmischke, M.: Interpretable approximation of high-dimensional data. SIAM J. Math. Data Sci. **3**(4), 1301–1323 (2021)
33. Potts, D., Schmischke, M.: Interpretable transformed ANOVA approximation on the example of the prevention of forest fires. Front. Appl. Math. Stat. **8** (2022)
34. Rahman, S.: Approximation errors in truncated dimensional decompositions. Math. Comput. **83**(290), 2799–2819 (2014)
35. Rahman, S.: A generalized ANOVA dimensional decomposition for dependent probability measures. SIAM-ASA J. Uncertain. **2**(1), 670–697 (2014)
36. Saha, E., Schaeffer, H., Tran, G.: HARFE: hard-ridge random feature expansion. Sampl. Theory Signal Process. Data Anal. **21**, 27 (2023)
37. Schmischke, M.: Dissertation: interpretable approximation of high-dimensional data based on the ANOVA decomposition. Universitaetsverlag Chemnitz (2022)
38. Sheather, S.J., Jones, M.C.: A reliable data-based bandwidth selection method for kernel density estimation. J. R. Stat. Soc. Ser. B Methodol. **53**, 683–690 (1991)
39. Sobol, I.M.: Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Math. Comput. Simul. **55**(1–3), 271–280 (2001)
40. Triebel, H.: Theory of function spaces III. Birkhäuser Basel, 1 edition, 01 (2006)
41. Wand, M., Jonas, M.: Kernel smoothing, vol. 60. London ; New York : Chapman & Hall (1995)
42. Weidensager, L., Krumm, D., Potts, D., Odenwald, S.: Estimating vertical ground reaction forces from plantar pressure using interpretable high-dimensional approximation. Sports Eng. (accepted) (2023)
43. Wu, C.F.J., Hamada, M.S.: Experiments - planning, analysis, and optimization. John Wiley & Sons, New York (2011)
44. Xie, Y., Shi, B., Schaeffer, H., Ward, R.: SHRIMP: sparser random feature models via iterative magnitude pruning. Math. Sci. Mach. Learn. PMLR **190**, 303–318 (2022)