Institut für Numerische Simulation
Rheinische Friedrich-Wilhelms-Universität Bonn
Wegelerstraße 6 • 53115 Bonn • Germany
phone +49 228 73-3427 • fax +49 228 73-7527
www.ins.uni-bonn.de

B. Bohn, M. Griebel

Error estimates for multivariate regression on discretized function spaces

INS Preprint No. 1412

March 2016

ERROR ESTIMATES FOR MULTIVARIATE REGRESSION ON DISCRETIZED FUNCTION SPACES

BASTIAN BOHN[†] AND MICHAEL GRIEBEL[†‡]

Abstract. In this paper, we will discuss the discretization error for the regression setting and derive error bounds relying on the approximation properties of the discretized space. Furthermore, we will point out how the sampling error and the discretization error interact and how they can be balanced appropriately. We will present two examples based on tensor product spaces (sparse grids, hyperbolic crosses) which provide a suitable approach in the case of large sample sets in moderate dimensions.

Key words. regression, learning theory, convergence rates, discretized function spaces, sparse grids, bias-variance problem

AMS subject classifications. 62J02 41A25 41A63

1. Introduction. Recent developments in mathematical learning theory [9],[18],[37] led to successful function regression algorithms and manifold learning methods such as support vector machines, kernel principal component analysis or principal manifold learning, see e.g. [19],[25],[31]. Here, error estimates for function regression as introduced in [8] rely on the approximation properties of the underlying function spaces. While there exist universal methods and error bounds for which the associated search set is not restricted to a subset of continuous functions, compare [2],[3],[24], most theory and algorithms rely on search sets that are subsets of reproducing kernel Hilbert spaces (RKHS). A thorough discussion of the error behaviour for regression in an RKHS can be found in [9].

The main advantage of reducing the search set to a subset of a reproducing kernel Hilbert space is the well-posedness of the underlying minimization problem. This is due to the compactness of the employed subsets in the space of continuous functions on a bounded domain. Furthermore, estimates for bounds on the sampling error in terms of covering numbers of compact search sets are available. Moreover, compactness is sufficient but not necessary for bounds on rates of convergence of the sampling error. Thus, there are also approaches based on non-compact function sets that fulfill the uniform Glivenko-Cantelli property, see e.g. [30],[40], which characterizes function sets with uniform sampling error convergence.

In a setting where we do not know the kernel function at all or where we do not have access to a closed form of the kernel function (e.g. for infinite series kernels), a discretized version of the RKHS must be used for the algorithm, see e.g. [17]. For the case of a large number of samples n a discretization of the search space is a way to get rid of the cubic costs with respect to n which standard kernel methods usually suffer from. Note that such a discretization does not necessarily lead to sample-dependent RKHS as in e.g. [39]. In our setting arbitrary discretizations of the kernel space can be used. Of course this introduces a certain discretization error which has to be controlled. This is in general not an easy task when norm-regularization of the corresponding minimization functional is considered.

In this paper, we will show how to account for the discretization error by using a-priori knowledge on the convergence rate of best approximation or interpolation errors with respect to the discretization level.[1] For specific settings there exist first results on the convergence

[†]Institute for Numerical Simulation, University of Bonn, Wegelerstr. 6, 53115 Bonn, Germany.

[‡]Fraunhofer Institute for Algorithms and Scientific Computing SCAI, Schloss Birlinghoven, 53754 Sankt Augustin, Germany

The authors were partially supported by the Sonderforschungsbereich 1060 *The Mathematics of Emergent Effects* funded by the Deutsche Forschungsgemeinschaft.

[1]Note that we do not compute the kernel of the discretized space to get sound error estimates for our approach.

behaviour when increasing the discretization level, see e.g. [12],[23],[27],[41]. More general approaches for squared ℓ_2 loss can be found in [7],[18],[32],[34]. Note however that specific convergence rates for the error have just been estimated for balls in infinite-dimensional Hilbert spaces or whole finite-dimensional spaces as search sets. Up to now only abstract results for balls in a finite-dimensional vector space can be found. Our results for the squared ℓ_2 loss-function can be seen as a specific instance of Theorem 3 from [7] with additional smoothness regularization or as versions of Theorem 4.1 from [32] and Theorem 1.5 from [34] with specific decay rates for the error. The more general concept of the distances between compact sets and L_2 is substituted in our version by the best approximation error in L_2 of finite-dimensional spaces V_h under the premise that the search set is a ball in V_h of sufficiently large radius. Furthermore, we relax the prerequisite on functions from the search set by using embedding norms instead of a fixed L_∞-bound (which would imply the so-called M-boundedness), see e.g. [9].

Our main result shows that the discretization error, measured in the so-called regression functional $\mathscr{E}$ to be minimized, can be estimated from above by the best approximation error in a certain L_2 Bochner space, i.e. we have

$$\mathscr{E}(f_{V_{h,b}}) - \mathscr{E}(\hat{f}) \leq C(b) \inf_{f \in V_h} \|f - \hat{f}\|_{L_{2,\rho_T}(T;\mathbb{R}^d)}$$

if b is appropriately coupled to the discretization level. Here, $f_{V_{h,b}}$ denotes the minimizer of $\mathscr{E}$ in the ball of radius b in the finite-dimensional space V_h and $\hat{f}$ denotes a minimizer over L_2. For the case of a squared ℓ_2 loss function in the error functional we even get a bound which depends quadratically on the best approximation error. This error estimate can easily be applied to discretized settings as in [4],[29]. There, sparse grid regression algorithms are introduced to deal with large sample sizes and circumvent the curse of dimensionality, i.e. the exponential growth of the computational costs with respect to the space dimension. Since no convergence rates have been derived for general sparse grid regression algorithms yet, we provide a detailed analysis of this case and show how to establish upper bounds for the overall convergence rate. As regression is closely connected to classification and density estimation, our general framework can, after a slight modification of the minimization functional, also be applied to discretized classification, see e.g. [21], and discretized density estimation, see e.g. [14],[38].

The outline of this paper is as follows: In section 2 we shortly review the minimization problem for (vector-valued) multivariate regression. Furthermore, we derive some useful properties of the error functional. In section 3 the splitting of the overall error into bias and sampling error (commonly used for regression, see e.g. [9]) is introduced. We discuss the relation of interpolation spaces and the bias for infinite-dimensional search sets in subsection 3.1. The case of finite-dimensional search sets and the induced discretization error is treated in subsection 3.2. Section 4 illustrates how an application of Hoeffding's inequality leads to an upper bound for the sampling error. In section 5 we comment on the specific settings of regression with piecewise linear ansatz functions and of regression with Fourier polynomials and apply our error bounds to the corresponding discretization spaces. There, we also discuss a proper balancing of the different error terms. Section 6 contains some concluding remarks.

2. Function regression. In this section, we introduce the general vector-valued function regression problem. Here, the measure underlying the data will be denoted by ρ. A short overview of the relevant functions and spaces/sets can be found in Table 1.

Let $T \subset \mathbb{R}^m$ be a compact domain (or manifold) and let μ be a measure on T. Further-

TABLE 1
Overview of relevant functions and spaces/sets for regression.

$L_{2,\rho_T}(T;\mathbb{R}^d)$	L_2 Bochner space with measure ρ_T
$\mathscr{H} \subset L_{2,\rho_T}(T;\mathbb{R}^d)$	infinite-dimensional Banach space, continuously embedded into $C(T;\mathbb{R}^d)$
$\mathscr{H}_b \subset \mathscr{H}$	closed ball of radius b in $\mathscr{H}$ centered at 0 w.r.t. the norm $\|\cdot\|_{\mathscr{H}}$
$V_h \subset L_{2,\rho_T}(T;\mathbb{R}^d)$	finite-dimensional Banach space, continuously embedded into $C(T;\mathbb{R}^d)$
$V_{h,b} \subset V_h$	closed ball of radius b in V_h centered at 0 w.r.t. the norm $\|\cdot\|_{V_h}$
$\hat{f} \in L_{2,\rho_T}(T;\mathbb{R}^d)$	minimizer of (2) in $L_{2,\rho_T}(T;\mathbb{R}^d)$
$f_W \in W \subset L_{2,\rho_T}(T;\mathbb{R}^d)$	minimizer of (2) in W
$f_{\mathscr{Z}_n,W} \in W \subset L_{2,\rho_T}(T;\mathbb{R}^d)$	minimizer of (9) in W

more, let

$$L_{p,\mu}(T;\mathbb{R}^d) := \left\{ f : T \to \mathbb{R}^d \mid \|f\|_{L_{p,\mu}(T;\mathbb{R}^d)} := \left(\int_T \|f(\mathbf{t})\|_{\ell_2}^p \mathrm{d}\mu(\mathbf{t}) \right)^{\frac{1}{p}} < \infty \right\} \tag{1}$$

denote the Bochner space of $\mathbb{R}^d$-valued functions for $1 \leq p \leq \infty$ with the usual modification

$$\|f\|_{L_{\infty,\mu}(T;\mathbb{R}^d)} := \inf\left\{ a \geq 0 \mid \mu(\{\mathbf{t} \in T \mid \|f(\mathbf{t})\|_{\ell_2} > a\}) = 0 \right\} < \infty$$

for the case $p = \infty$. A general multivariate function regression problem then reads as follows:

$$\text{Find } \hat{f} := \operatorname*{arg\,min}_{f \in L_{2,\rho_T}(T;\mathbb{R}^d)} \mathscr{E}(f) \text{ with } \mathscr{E}(f) := \int_{T \times \mathbb{R}^d} \psi(f(\mathbf{t}),\mathbf{x}) \, \mathrm{d}\rho(\mathbf{t},\mathbf{x}). \tag{2}$$

Here, ρ is a probability measure on the Borel algebra of $T \times \mathbb{R}^d$ for which the marginal measure

$$\rho_T(\cdot) := \rho(\cdot, \mathbb{R}^d)$$

with respect to T and the conditional measure

$$\rho(\mathbf{x}|\mathbf{t}) := \frac{\rho(\mathbf{t},\mathbf{x})}{\rho_T(\mathbf{t})}$$

with respect to $\mathbf{t}$ are also probability measures. Furthermore, $\psi : \mathbb{R}^d \times \mathbb{R}^d \to [0,\infty)$ is a fixed cost function which penalizes large distances between the two input vectors. Therefore, for a specific $f \in L_{2,\rho_T}(T;\mathbb{R}^d)$, the value $\psi(f(\mathbf{t}),\mathbf{x})$ can be interpreted as the error which is made by using f to generate $\mathbf{x}$ from a given $\mathbf{t}$. Integrating over $T \times \mathbb{R}^d$ we obtain $\mathscr{E}$, which resembles the average error modelled by ψ. The measure ρ resembles the significance of the data, i.e. how likely it is that a specific $\mathbf{t}$ should be matched to a specific $\mathbf{x}$. Since we are looking for a function which best describes the interplay between $\mathbf{t}$ and $\mathbf{x}$ (drawn according to ρ) on average, (2) suits our purpose. Throughout this paper we will be interested in estimating the error $\mathscr{E}(f) - \mathscr{E}(\hat{f})$ for a given function $f \in L_{2,\rho_T}(T;\mathbb{R}^d)$ which is either analytically defined or has been computed by an algorithm. Therefore, the difference[2] $\mathscr{E}(f) - \mathscr{E}(\hat{f})$ indicates how close $\mathscr{E}(f)$ is to the true minimum $\mathscr{E}(\hat{f})$.

[2] In other settings, e.g. if one estimates the density ρ directly, another error measure such as the Kullback-Leibler divergence or a other divergences might be more appropriate here.

For the remainder of this paper we assume that

$$\psi(\mathbf{x},\mathbf{y}) = \tilde{\psi}(\mathbf{x}-\mathbf{y}) \tag{3}$$

for an even and convex function $\tilde{\psi}$, which fulfills the following weakened Lipschitz condition: For each $\tilde{M} > 0$, there exists a $C > 0$ such that

$$|\tilde{\psi}(\mathbf{x}_1) - \tilde{\psi}(\mathbf{x}_2)| \leq C\|\mathbf{x}_1 - \mathbf{x}_2\|_{\ell_2}. \tag{4}$$

for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$ with $\max\left(\|\mathbf{x}_1\|_{\ell_2}, \|\mathbf{x}_2\|_{\ell_2}\right) \leq \tilde{M}$. The so-called ε-insensitive loss function

$$\psi(\mathbf{x},\mathbf{y}) = \begin{cases} \|\mathbf{x}-\mathbf{y}\|_{\ell_2} - \varepsilon & \text{if } \|\mathbf{x}-\mathbf{y}\|_{\ell_2} > \varepsilon \\ 0 & \text{else} \end{cases}$$

is an example for a cost function which fulfills these requirements and is often used in machine learning, see e.g. [31, 37]. While most of our results hold for a much larger class of cost functions, the Lipschitz-property is needed to establish the results of section 3.

Another widely used cost function is the squared norm

$$\psi(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_{\ell_2}^2, \tag{5}$$

which also fulfills the above requirements since

$$\begin{aligned} \|\mathbf{x}_1\|_{\ell_2}^2 - \|\mathbf{x}_2\|_{\ell_2}^2 &= \left(\|\mathbf{x}_1\|_{\ell_2} + \|\mathbf{x}_2\|_{\ell_2}\right)\left(\|\mathbf{x}_1\|_{\ell_2} - \|\mathbf{x}_2\|_{\ell_2}\right) \\ &\leq 2\tilde{M}\left(\|\mathbf{x}_1\|_{\ell_2} - \|\mathbf{x}_2\|_{\ell_2}\right) \leq 2\tilde{M}\|\mathbf{x}_1 - \mathbf{x}_2\|_{\ell_2} \end{aligned}$$

holds if $\max\left(\|\mathbf{x}_1\|_{\ell_2}, \|\mathbf{x}_2\|_{\ell_2}\right) \leq \tilde{M}$. Thus, $C = 2\tilde{M}$ is a valid choice in (4).

We assume[3] that there exists an $r > 0$ such that

$$\rho(T \times \overline{U_r^d(\mathbf{0})}) = 1 \tag{6}$$

where $U_r^d(\mathbf{0})$ denotes the open ℓ_2-ball of radius r with center $\mathbf{0}$ in $\mathbb{R}^d$. Then, it can easily be seen that

$$f_\rho(\cdot) := \int_{\mathbb{R}^d} \mathbf{x}\, \mathrm{d}\rho(\mathbf{x}|\cdot) \in L_{\infty,\rho_T}(T;\mathbb{R}^d) \subset L_{2,\rho_T}(T;\mathbb{R}^d). \tag{7}$$

Therefore, for the choice (5), f_ρ is a solution to (2) and we have $\hat{f} = f_\rho$.

In real-world applications the measure ρ is however unknown and we only have access to a set $\mathscr{Z}_n$ of finitely many sample points

$$\mathscr{Z}_n := ((\mathbf{t}_i, \mathbf{x}_i))_{i=1}^n \in \left(T \times \mathbb{R}^d\right)^n \tag{8}$$

which we assume to be drawn independently and to be distributed according to ρ. Let us introduce the new measures

$$\delta_{\mathscr{Z}_n} := \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{t}_i,\mathbf{x}_i} \quad \text{and} \quad \delta_{\mathbf{t}} := \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{t}_i}$$

[3] Note that this condition can be relaxed to

$$\int_{\mathbb{R}^d} \|\mathbf{x}\|_{\ell_2}\, \mathrm{d}\rho(\mathbf{x}|\mathbf{t}) \leq r \;\forall\, \mathbf{t} \in T$$

if we assume the so-called M-boundedness, which is introduced in section 4, see e.g. [9] for this case.

on $T \times \mathbb{R}^d$ and T, respectively. Here $\delta_{\mathbf{t}_i,\mathbf{x}_i}$ is the usual Dirac measure centered in $(\mathbf{t}_i,\mathbf{x}_i) \in T \times \mathbb{R}^d$ and $\delta_{\mathbf{t}_i}$ is the Dirac measure centered in $\mathbf{t}_i \in T$. Then, substituting ρ by $\delta_{\mathscr{Z}_n}$, the regression problem for a finite sample set (8) reads as follows:

$$\text{Find} \quad \operatorname*{arg\,min}_{f \in L_{2,\rho_T}(T;\mathbb{R}^d)} \mathscr{E}_{\mathscr{Z}_n}(f) \quad \text{where } \mathscr{E}_{\mathscr{Z}_n}(f) := \frac{1}{n}\sum_{i=1}^{n} \psi(f(\mathbf{t}_i),\mathbf{x}_i). \tag{9}$$

In the following, we will rely on the Lipschitz continuity of the functionals $\mathscr{E}$ and $\mathscr{E}_{\mathscr{Z}_n}$ which relies on the Lipschitz continuity of $\tilde{\psi}$.

LEMMA 1. *Let f_1, f_2 be such that $\|f_i\|_{L_{\infty,\rho_T}(T;\mathbb{R}^d)} \le M$ for $i = 1,2$ and let $C > 0$ be a constant such that (4) holds for the norm bound $\tilde{M} := M + r$ with r from (6). Then*

$$|\mathscr{E}(f_1) - \mathscr{E}(f_2)| \le C\|f_1 - f_2\|_{L_{1,\rho_T}(T;\mathbb{R}^d)} \le C\|f_1 - f_2\|_{L_{2,\rho_T}(T;\mathbb{R}^d)} \tag{10}$$

and

$$|\mathscr{E}_{\mathscr{Z}_n}(f_1) - \mathscr{E}_{\mathscr{Z}_n}(f_2)| \le C\|f_1 - f_2\|_{L_{1,\delta_{\mathbf{t}}}(T;\mathbb{R}^d)} \le C\|f_1 - f_2\|_{L_{2,\delta_{\mathbf{t}}}(T;\mathbb{R}^d)}. \tag{11}$$

Proof. Since $\max\left(\|f_1(\mathbf{t}) - \mathbf{x}\|_{\ell_2}, \|f_2(\mathbf{t}) - \mathbf{x}\|_{\ell_2}\right) \le M + r$ for $i = 1,2$ and ρ-almost every $\mathbf{t}$ and $\mathbf{x}$, we have

$$\begin{aligned}
|\mathscr{E}(f_1) - \mathscr{E}(f_2)| &\le \int_{T\times\mathbb{R}^d} |\psi(f_1(\mathbf{t}),\mathbf{x}) - \psi(f_2(\mathbf{t}),\mathbf{x})| \,\mathrm{d}\rho(\mathbf{t},\mathbf{x}) \\
&\le C \int_{T\times\mathbb{R}^d} \|f_1(\mathbf{t}) - f_2(\mathbf{t})\|_{\ell_2} \mathrm{d}\rho(\mathbf{t},\mathbf{x}) \\
&= C \int_T \|f_1(\mathbf{t}) - f_2(\mathbf{t})\|_{\ell_2} \int_{\mathbb{R}^d} \mathrm{d}\rho(\mathbf{x}|\mathbf{t}) \,\mathrm{d}\rho_T(\mathbf{t}) \\
&= C \int_T \|f_1(\mathbf{t}) - f_2(\mathbf{t})\|_{\ell_2} \,\mathrm{d}\rho_T(\mathbf{t}) \\
&= C\|f_1 - f_2\|_{L_{1,\rho_T}(T;\mathbb{R}^d)}.
\end{aligned}$$

Jensen's inequality then yields $\|f_1 - f_2\|_{L_{1,\rho_T}(T;\mathbb{R}^d)} \le \|f_1 - f_2\|_{L_{2,\rho_T}(T;\mathbb{R}^d)}$ for all $f_1 - f_2 \in L_{1,\rho_T}(T;\mathbb{R}^d)$ which finally shows (10).

The proof for the inequality (11) works analogously: We have

$$\begin{aligned}
|\mathscr{E}_{\mathscr{Z}_n}(f_1) - \mathscr{E}_{\mathscr{Z}_n}(f_2)| &\le \frac{1}{n}\sum_{i=1}^{n} |\psi(f_1(\mathbf{t}_i),\mathbf{x}_i) - \psi(f_2(\mathbf{t}_i),\mathbf{x}_i)| \\
&\le \frac{C}{n}\sum_{i=1}^{n} \|f_1(\mathbf{t}_i) - f_2(\mathbf{t}_i)\|_{\ell_2} = C\|f_1 - f_2\|_{L_{1,\delta_{\mathbf{t}}}(T;\mathbb{R}^d)}.
\end{aligned}$$

Applying Jensen's inequality to $\|f_1 - f_2\|_{L_{1,\delta_{\mathbf{t}}}(T;\mathbb{R}^d)}$ gives the final result (11). ∎

In most papers, the bound M - and thus also $\tilde{M}$ - is assumed to be fix. We will here discuss both, the case of fixed M and the case where we only rely on norm bounds in the corresponding search set instead. We suggest the second approach to reflect the situation in specific regression algorithms more appropriately, since there a fixed bound on the L_∞ norm is usually not present in the first place.

We now derive another representation of $\mathscr{E}(f) - \mathscr{E}(f_\rho)$ for the specific cost function (5).

LEMMA 2. *Let $f \in L_{2,\rho_T}(T;\mathbb{R}^d)$ and let $\psi(\mathbf{x},\mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_{\ell_2}^2$, cf. (5). Then*

$$|\mathscr{E}(f) - \mathscr{E}(f_\rho)| = \|f - f_\rho\|_{L_{2,\rho_T}(T;\mathbb{R}^d)}^2. \tag{12}$$

Proof. It holds

$$\begin{aligned}\mathscr{E}(f) &= \int_{T\times\mathbb{R}^d} \|f(\mathbf{t}) - f_\rho(\mathbf{t}) + f_\rho(\mathbf{t}) - \mathbf{x}\|_{\ell_2}^2 \,\mathrm{d}\rho(\mathbf{t},\mathbf{x}) \\ &= \int_{T\times\mathbb{R}^d} \|f(\mathbf{t}) - f_\rho(\mathbf{t})\|_{\ell_2}^2 \,\mathrm{d}\rho(\mathbf{t},\mathbf{x}) + \int_{T\times\mathbb{R}^d} \|f_\rho(\mathbf{t}) - \mathbf{x}\|_{\ell_2}^2 \,\mathrm{d}\rho(\mathbf{t},\mathbf{x}) \\ &+ \int_{T\times\mathbb{R}^d} 2\langle f(\mathbf{t}) - f_\rho(\mathbf{t}), f_\rho(\mathbf{t}) - \mathbf{x}\rangle \,\mathrm{d}\rho(\mathbf{t},\mathbf{x}) \\ &= \|f - f_\rho\|_{L_{2,\rho_T}(T;\mathbb{R}^d)}^2 + \mathscr{E}(f_\rho) + 2\int_{T\times\mathbb{R}^d} \langle f(\mathbf{t}) - f_\rho(\mathbf{t}), f_\rho(\mathbf{t}) - \mathbf{x}\rangle \,\mathrm{d}\rho(\mathbf{t},\mathbf{x}).\end{aligned}$$

Here, $\langle\cdot,\cdot\rangle$ denotes the standard scalar product in $\mathbb{R}^d$. Note that the last summand is zero because of the definition of f_ρ. Therefore (12) is proven. ∎

3. The Bias. We now further decompose the overall error into its bias error part and its sampling error part.

The original task was to solve the minimization problem (2) for $f \in L_{2,\rho_T}(T;\mathbb{R}^d)$. However, since we deal with point evaluations of functions in (9), it makes sense to restrict our search to a space which is continuously embedded into the space $C(T;\mathbb{R}^d)$ of vector-valued continuous functions equipped with the maximum norm $\|f\|_\infty := \sup_{\mathbf{t}\in T} \|f(t)\|_{\ell_2}$. For an actual implementation of a minimization algorithm, the search set is then often restricted to a bounded ball in this search space. To this end, we have to distinguish two cases: First, we may consider an infinite-dimensional search space $\mathscr{H}$ which is dense in $L_{2,\rho_T}(T;\mathbb{R}^d)$. For this case some results on the bias are already known in the literature. For the sake of completeness, we will recap them in subsection 3.1. Then, in subsection 3.2 we consider the case of a finite-dimensional search space V_h. Note here that the known results from subsection 3.1 cannot be applied. We therefore will derive new results on the bias in this case.

3.1. Infinite-dimensional search spaces. Most concepts introduced in this section follow [8],[9]. We assume that $(\mathscr{H}, \|\cdot\|_\mathscr{H})$ is an infinite-dimensional Banach space which is dense in $L_{2,\rho_T}(T;\mathbb{R}^d)$ and fulfills the relation

$$\mathscr{H} \subset C(T;\mathbb{R}^d) \subset L_{\infty,\rho_T}(T;\mathbb{R}^d) \subset L_{2,\rho_T}(T;\mathbb{R}^d). \tag{13}$$

Furthermore, we assume that all embeddings

$$\begin{aligned}(\mathscr{H}, \|\cdot\|_\mathscr{H}) &\hookrightarrow (C(T;\mathbb{R}^d), \|\cdot\|_\infty) \\ &\hookrightarrow (L_{\infty,\rho_T}(T;\mathbb{R}^d), \|\cdot\|_{L_{\infty,\rho_T}(T;\mathbb{R}^d)}) \hookrightarrow (L_{2,\rho_T}(T;\mathbb{R}^d), \|\cdot\|_{L_{2,\rho_T}(T;\mathbb{R}^d)})\end{aligned} \tag{14}$$

are continuous.[4] In addition to reproducing kernel Hilbert spaces, we can here also incorporate certain Banach spaces, e.g. Sobolev spaces with respect to the L_p norm for $p \neq 2$. More information on reproducing kernel Hilbert spaces of vector-valued functions and operator-valued kernels can be found in [26].

[4] Note that this assumption is not only fulfilled by reproducing kernel Hilbert spaces, but also by any space which is continuously embedded into an RKHS. To this end, let $\mathscr{G}$ be an RKHS with norm $\|\cdot\|_\mathscr{G}$ and let the embedding $(\mathscr{H}, \|\cdot\|_\mathscr{H}) \hookrightarrow (\mathscr{G}, \|\cdot\|_\mathscr{G})$ be continuous. Then, for any probability measure μ it holds

$$\|g\|_{L_{2,\mu}(T;\mathbb{R}^d)} \leq \|g\|_{L_{\infty,\mu}(T;\mathbb{R}^d)} \leq \|g\|_\infty \leq \sup_{\mathbf{t}\in T} \sqrt{\|K(\mathbf{t},\mathbf{t})\|_2}\|g\|_\mathscr{G} \leq c_{\mathscr{H},\mathscr{G}} \sup_{\mathbf{t}\in T} \sqrt{\|K(\mathbf{t},\mathbf{t})\|_2}\|g\|_\mathscr{H}$$

for all $g \in \mathscr{H}$. Here, $c_{\mathscr{H},\mathscr{G}}$ denotes the operator norm of the continuous embedding $\mathrm{id} : \mathscr{H} \hookrightarrow \mathscr{G}$, $K : T \times T \to \mathscr{L}(\mathbb{R}^d,\mathbb{R}^d)$ denotes the reproducing kernel of $\mathscr{G}$ and $\|\cdot\|_2$ is the operator norm for matrices in $\mathbb{R}^{d\times d}$.

We now employ the bounded ball

$$\mathscr{H}_b := \{f \in \mathscr{H} \mid \|f\|_{\mathscr{H}} \leq b\}$$

as our actual search set. Thus, we encounter the specific situation

$$\mathscr{H}_b \subset \mathscr{H} \subset C(T;\mathbb{R}^d) \subset L_{\infty,\rho_T}(T;\mathbb{R}^d) \subset L_{2,\rho_T}(T;\mathbb{R}^d), \tag{15}$$

compare (13).[5] From now on, we assume that $\mathscr{H}_b$ is a compact subset of $C(T;\mathbb{R}^d)$. Using the Arzela-Ascoli Theorem, one can show that this assumption is fulfilled for a reproducing kernel Hilbert space, see e.g. section 2.6 of [9]. For a continuously embedded subspace $\mathscr{H}$ of an RKHS we obtain the compactness of $\mathscr{H}_b$ if it is closed with respect to $\|\cdot\|_\infty$. The restriction to $\mathscr{H}_b$ corresponds to a specific regularization which eliminates the ill-posedness[6] of the original problems (2) and (9).

PROPOSITION 3. *Let $b > 0$. The minimizers to (2) and (9) exist if we restrict the search set to $\mathscr{H}_b$.*

Proof. Note that

$$\|f\|_{L_{\infty,\rho_T}(T;\mathbb{R}^d)} \leq c_{\mathscr{H},C(T;\mathbb{R}^d)}\|f\|_{\mathscr{H}} \leq b \cdot c_{\mathscr{H},C(T;\mathbb{R}^d)}$$

for all $f \in \mathscr{H}_b$, where $c_{\mathscr{H},C(T;\mathbb{R}^d)}$ is the corresponding embedding constant. Therefore, the error functionals $\mathscr{E}$ and $\mathscr{E}_{\mathscr{Z}_n}$ are Lipschitz-continuous on $\mathscr{H}_b$ with respect to the $L_{2,\rho_T}(T;\mathbb{R}^d)$ norm or the $L_{2,\delta_{\mathfrak{t}}}(T;\mathbb{R}^d)$ norm, respectively, see Lemma 1. Thus, they are also Lipschitz-continuous on $C(T;\mathbb{R}^d)$. Due to the compactness of $\mathscr{H}_b$ in $C(T;\mathbb{R}^d)$ for every $b > 0$ we obtain existence of the minimizers of $\mathscr{E}$ over $\mathscr{H}_b$ and $\mathscr{E}_{\mathscr{Z}_n}$ over $\mathscr{H}_b$. □

Here, $b > 0$ is a free parameter. To fix our notation, let $f_{\mathscr{H}_b}$ be a minimizer of (2) where the search set is restricted to $\mathscr{H}_b$. Then, we can write the so-called *bias* as

$$\mathscr{E}(f_{\mathscr{H}_b}) - \mathscr{E}(\hat{f}) \tag{16}$$

which is a measure for the error that occurs due to the restriction of the search set from $L_{2,\rho_T}(T;\mathbb{R}^d)$ to $\mathscr{H}_b$.

Next, let us define the interpolation space $(X,Y)_\zeta$ of a Banach space X and a subspace $Y \subset X$ for $\zeta \in (0,1)$. Following [28], this is the space which consists of all functions $g \in X$ with

$$\|g\|_\zeta := \sup_{t>0} \frac{\mathbb{K}(g,t)}{t^\zeta} < \infty,$$

where the so-called $\mathbb{K}$-functional is defined by

$$\mathbb{K}(g,t) := \inf_{h\in Y}(\|g-h\|_X + t\|h\|_Y).$$

Observe here that the value of the $\mathbb{K}$-functional approaches the best approximation error $\inf_{h\in Y}\|g-h\|_X$ for small t. Therefore, the norm $\|\cdot\|_\zeta$ expresses the decay properties of this best approximation error.

As the embedding $\mathscr{H} \hookrightarrow L_{2,\rho_T}(T;\mathbb{R}^d)$ is continuous, see (14), we have the following result: Error bounds for the bias are determined by the largest $1 > \zeta > 0$ such that $\hat{f}$ is an element of the interpolation space $\left(L_{2,\rho_T}(T;\mathbb{R}^d),\mathscr{H}\right)_\zeta$, see [9]. More precisely, the following theorem holds.

[5] There also exist results on the minimization in the L_{2,ρ_T}-setting, see e.g. [2],[24]. We restrict ourselves here to the more practical situation where the search set is a compact (in $C(T;\mathbb{R}^d)$) ball in $\mathscr{H}$.

[6] If $\hat{f} \in L_{2,\rho_T}(T;\mathbb{R}^d)$, e.g. for the cost function (5), the problem (2) is already well-posed and only the discrete sample problem (9) has to be regularized.

THEOREM 4. *Let $\hat{f} \in L_{\infty,\rho_T}(T;\mathbb{R}^d)$ be a minimizer of (2) and let ψ be a cost function as in (3). Let furthermore C be the constant from Lemma 1 for $M = \max_{f\in\mathscr{H}_b\cup\{\hat{f}\}} \|f\|_{L_{\infty,\rho_T}(T;\mathbb{R}^d)}$. If $\hat{f} \in \left(L_{2,\rho_T}(T;\mathbb{R}^d),\mathscr{H}\right)_{\frac{\theta}{\theta+2}}$ for a $\theta \in (0,\infty)$, we obtain*

$$\mathscr{E}(f_{\mathscr{H}_b}) - \mathscr{E}(\hat{f}) \le C\|\hat{f}\|^{\frac{\theta+2}{2}}_{\frac{\theta}{\theta+2}} \cdot b^{-\frac{\theta}{2}} \tag{17}$$

Furthermore, for the cost function $\psi(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|^2_{\ell_2}$, see (5), we have the following equivalence:

$$\hat{f} \in \left(L_{2,\rho_T}(T;\mathbb{R}^d),\mathscr{H}\right)_{\frac{\theta}{\theta+2}} \quad\Leftrightarrow\quad \mathscr{E}(f_{\mathscr{H}_b}) - \mathscr{E}(\hat{f}) \le \|\hat{f}\|^{\theta+2}_{\frac{\theta}{\theta+2}} \cdot b^{-\theta} < \infty. \tag{18}$$

Proof. The statement (17) follows directly from Lemma 1 and Theorem 4.16 of [9]. In the special case $\psi(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|^2_{\ell_2}$ we have $\hat{f} = f_\rho$ and the proof of (18) follows from Lemma 2 and Theorem 4.16 of [9]. ∎

For a reproducing kernel Hilbert space, the convergence rate of the bias (for $b\to\infty$) is therefore mainly governed by the decay of the eigenvalues of the kernel's integral operator. For more details, we refer to [10] for an introduction to interpolation spaces and to [9] for details on the connection between these spaces and the bias.

Note that C in Theorem 4 might depend on b. Thus, assuming that $C \sim b^\kappa$ for some $\kappa > 0$, the exponent of the rate in (17) becomes $-\frac{\theta}{2}+\kappa$ compared to $-\theta$ as in (18) for the squared ℓ_2 costs.

3.2. The discretization error. To be able to solve the minimization problem (9) over $\mathscr{H}_b$ on a computer, we consider a further restriction of the search space to a finite-dimensional space V_h. This can be understood as solving a different, finite-dimensional problem instead of considering a minimization over $\mathscr{H}_b$. Alternatively, we could consider V_h as a discretization to $\mathscr{H}$. Here, the subscript h refers to the mesh-width or, more generally, to the approximation properties of the finite-dimensional space. Then, as we will point out later in this section, we cannot rely on the bounds for the bias we derived in section 3.1 when dealing with finite-dimensional search spaces.

To be precise, similarly as in the infinite-dimensional case, V_h is a (finite-dimensional) normed subspace of $C(T;\mathbb{R}^d)$. Let us denote a ball in V_h by

$$V_{h,b} = \{f \in V_h \mid \|f\|_{V_h} \le b\}.$$

Here, $\|\cdot\|_{V_h}$ serves as regularization norm,[7] which is allowed to depend on the discretization parameter h. We assume that the embedding $V_h \hookrightarrow C(T;\mathbb{R}^d)$ is continuous with embedding constant

$$c_{V_h} := \|\mathrm{id} : V_h \hookrightarrow C(T;\mathbb{R}^d)\|_{\mathscr{L}(V_h,C(T;\mathbb{R}^d))},$$

where $\|\cdot\|_{\mathscr{L}(X,Y)}$ is the standard norm for linear operators from X to Y. Thus $V_{h,b}$ is a compact subset of $C(T;\mathbb{R}^d)$ and we have

$$V_{h,b} \subset V_h \subset C(T;\mathbb{R}^d) \subset L_{\infty,\rho_T}(T;\mathbb{R}^d) \subset L_{2,\rho_T}(T;\mathbb{R}^d).$$

The restriction to the set $V_{h,b}$ in the finite-dimensional space V_h introduces an associated bias, see section 3.1. As we now deal with finite-dimensional search spaces, we call this bias

[7] Later on, for the examples in this paper, we will only use the Sobolev norm of mixed smoothness of order 1, i.e. $\|\cdot\|_{V_h} = \|\cdot\|_{H^1_{\mathrm{mix}}}$, which is independent of h anyway.

"discretization error". A straightforward way to account for this discretization error would be to apply results from interpolation theory. Let us discuss this approach shortly: Since $\dim(V_h) < \infty$, the space V_h corresponding to the search set $V_{h,b}$ is no longer dense in L_{2,ρ_T}. Thus, the known results from Theorem 4 are no longer useful. This is due to the fact that the $\mathbb{K}$-functional for the pair $\left(L_{2,\rho_T}, V_h\right)$ cannot approach zero for $t \to 0$ if the considered function $\hat{f}$ is not an element of V_h. Therefore, $\hat{f}$ is not an element of the interpolation space $\left(L_{2,\rho_T}, V_h\right)_{\frac{\theta}{2+\theta}}$ for any $\theta > 0$ and we cannot derive a convergence rate for this new bias with the conventional methods from section 3.1.

To deal with our discretization error we could employ a result from [32]. There, it was shown that for certain sequences of linear operators $\mathscr{L}_n : L_{2,\rho_T} \to \mathscr{H}$, $n > 0$, for which specific Jackson and Bernstein inequalities hold, there exists $n_0 > 0$ such that the error $\hat{f} - \mathscr{L}_n \hat{f}$ is dominated by the bias $\mathscr{E}(f_{\mathscr{H}_b}) - \mathscr{E}(\hat{f})$ for all $n > n_0$. This can also be used to incorporate interpolation or best approximation operators $\mathscr{L}_n$ mapping into $V_h \subset \mathscr{H}$. Although this result is very general, the spaces V_h have to be a subset of the space $\mathscr{H}$, i.e. the space for which a Jackson inequality is provided. In other words, this means that the discretization needs to be conforming. But since we want to apply our method to e.g. spline spaces on sparse grids, which are not conforming due to the fact that $V_h \nsubseteq H^2_{\text{mix}}$ for a piecewise linear spline space V_h, we cannot apply the results from [32] directly. The definition of H^2_{mix} and a more detailed discussion on this subject can be found in section 5.1. We will therefore introduce a related technique which also relies on Jackson and Bernstein inequalities of certain best approximation operators but which is not restricted to conforming discretization spaces.

To this end, let us denote the minimizer of (2) over $V_{h,b}$ by $f_{V_{h,b}}$ and let us denote the minimizer of (9) over $V_{h,b}$ by $f_{\mathscr{Z}_n,V_{h,b}}$. Note that the existence of these minimizers follows from Proposition 3 by substituting $\mathscr{H}$ by V_h there. As usual, the overall error is decomposed as

$$\begin{aligned} \mathscr{E}(f_{\mathscr{Z}_n,V_{h,b}}) - \mathscr{E}(\hat{f}) = \mathscr{E}(f_{V_{h,b}}) - \mathscr{E}(\hat{f}) \\ + \mathscr{E}(f_{\mathscr{Z}_n,V_{h,b}}) - \mathscr{E}(f_{V_{h,b}}), \end{aligned} \tag{19}$$

where the first part is the bias/discretization error and the second part describes the sampling error in the finite-dimensional setting which we deal with in section 4. The basic idea of our approach is to choose the norm bound b such that

$$\inf_{f \in V_{h,b}} \|f - \hat{f}\|_{L_{2,\rho_T}(T;\mathbb{R}^d)} = \inf_{f \in V_h} \|f - \hat{f}\|_{L_{2,\rho_T}(T;\mathbb{R}^d)}. \tag{20}$$

Then, the first part $\mathscr{E}(f_{V_{h,b}}) - \mathscr{E}(\hat{f})$ of (19) can be bounded by the L_2-best approximation error in V_h. Now, the norm bound b influences just the upper bound for the second part $\mathscr{E}(f_{\mathscr{Z}_n,V_{h,b}}) - \mathscr{E}(f_{V_{h,b}})$ of (19). As we will see in section 4, this term grows if b increases. Therefore, we will choose b as small as possible such that (20) is fulfilled.

We will first prove a Lemma that shows how b has to be chosen with respect to the L_2 norm of $\hat{f}$.

LEMMA 5. *Let V_h be such that the inverse (Bernstein) inequality*

$$\|f\|_{V_h} \leq c(h) \, \|f\|_{L_{2,\rho_T}(T;\mathbb{R}^d)} \tag{21}$$

holds for every $f \in V_h$. Then, the best $L_{2,\rho_T}(T;\mathbb{R}^d)$ approximation f^{BA}_{h,ρ_T} from V_h to the function $\hat{f}$ fulfills

$$\|f^{BA}_{h,\rho_T}\|_{V_h} \leq c(h) \cdot \|\hat{f}\|_{L_{2,\rho_T}(T;\mathbb{R}^d)}$$

Therefore, $f^{BA}_{h,\rho_T} \in V_{h,b}$ *for*

$$b := c(h) \cdot \|\hat{f}\|_{L_{2,\rho_T}(T;\mathbb{R}^d)}. \tag{22}$$

Proof. In the following, we write L_2 for $L_{2,\rho_T}(T;\mathbb{R}^d)$ to simplify notation. Note that $\langle \hat{f} - f^{\mathrm{BA}}_{h,\rho_T}, f^{\mathrm{BA}}_{h,\rho_T} \rangle_{L_2} = 0$ since $\mathrm{Id} - P$ is orthogonal on V_h for the orthogonal projector $P : L_2 \to V_h$. Therefore, it holds

$$\|\hat{f}\|^2_{L_2} = \|\hat{f} - f^{\mathrm{BA}}_{h,\rho_T} + f^{\mathrm{BA}}_{h,\rho_T}\|^2_{L_2} = \|\hat{f} - f^{\mathrm{BA}}_{h,\rho_T}\|^2_{L_2} + \|f^{\mathrm{BA}}_{h,\rho_T}\|^2_{L_2} \geq \|f^{\mathrm{BA}}_{h,\rho_T}\|^2_{L_2}$$

and we get

$$\|f^{\mathrm{BA}}_{h,\rho_T}\|_{L_2} \leq \|\hat{f}\|_{L_2}. \tag{23}$$

Thus, since $f^{\mathrm{BA}}_{h,\rho_T} \in V_h$, we have with (21)

$$\|f^{\mathrm{BA}}_{h,\rho_T}\|_{V_h} \leq c(h)\|f^{\mathrm{BA}}_{h,\rho_T}\|_{L_2} \leq c(h)\|\hat{f}\|_{L_2}. \qquad \square$$

We are now in the position to establish a bound for the discretization error.

THEOREM 6. *Let* $\hat{f} \in L_{\infty,\rho_T}(T;\mathbb{R}^d)$ *solve* (2)*, let* b *be chosen as in Lemma 5 and let* ψ *be a cost function as in* (3)*. Let furthermore* C *fulfill the prerequisite of Lemma 1 for* $M = \max_{f \in V_{h,b} \cup \{\hat{f}\}} \|f\|_{L_{\infty,\rho_T}(T;R^d)}$*. Then, the discretization error can be bounded by*

$$\mathscr{E}(f_{V_{h,b}}) - \mathscr{E}(\hat{f}) \leq C \inf_{f \in V_h} \|f - \hat{f}\|_{L_{2,\rho_T}(T;\mathbb{R}^d)}. \tag{24}$$

Furthermore, if $\psi(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|^2_{\ell_2}$ *we even have*

$$\mathscr{E}(f_{V_{h,b}}) - \mathscr{E}(\hat{f}) = \inf_{f \in V_h} \|f - \hat{f}\|^2_{L_{2,\rho_T}(T;\mathbb{R}^d)}.$$

Proof. It holds

$$\begin{aligned}
\mathscr{E}(f_{V_{h,b}}) - \mathscr{E}(\hat{f}) &\leq \inf_{f \in V_{h,b}} |\mathscr{E}(f) - \mathscr{E}(\hat{f})| \\
&\overset{\text{Lemma 1}}{\leq} C \inf_{f \in V_{h,b}} \|f - \hat{f}\|_{L_{2,\rho_T}(T;\mathbb{R}^d)} \\
&\overset{\text{Lemma 5}}{=} C \inf_{f \in V_h} \|f - \hat{f}\|_{L_{2,\rho_T}(T;\mathbb{R}^d)},
\end{aligned}$$

which proves (24). For the special case $\psi(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|^2_{\ell_2}$ we have $\hat{f} = f_\rho$ and thus

$$\begin{aligned}
\mathscr{E}(f_{V_{h,b}}) - \mathscr{E}(\hat{f}) &\overset{\text{Lemma 2}}{=} \inf_{f \in V_{h,b}} \|f - f_\rho\|^2_{L_{2,\rho_T}(T;\mathbb{R}^d)} \\
&\overset{\text{Lemma 5}}{=} \inf_{f \in V_h} \|f - f_\rho\|^2_{L_{2,\rho_T}(T;\mathbb{R}^d)}. \qquad \square
\end{aligned}$$

Note that the prerequisite $\hat{f} \in L_{\infty,\rho_T}(T;\mathbb{R}^d)$ in Theorem 6 is not a severe restriction at all since $\hat{f}$ approximates the data which is almost surely bounded by $\|\mathbf{x}\|_{\ell_2} \leq r$ anyway, see (6). Altogether, we have obtained a bound (24) for the discretization error which is determined by the error of the best approximation in the discretized space V_h and the constant C from Lemma 1. When estimating the latter, note that we deal in Theorem 6 with functions from

$V_{h,b}$. Therefore, if we do not assume C to be an universal constant the growth of C will be governed by the behaviour of b. The price we paid to achieve the error bounds of Theorem 6 is the coupling (22) of the norm-bound b to the Bernstein-factor $c(h)$.

From our results we see that we obtain a small error bound for the discretization error if the L_2-best approximation error to $\hat{f}$ is small for functions in V_h. Therefore, the smoothness of $\hat{f}$ has to be exploited by a suitable Jackson inequality. On the other hand, the discretized sampling error becomes small for small b and small spaces V_h. These relations have to be taken into account when balancing the two error terms later on.

4. The sampling error. In this section we consider the sampling error in more detail. Here, we follow [9]. Although we deal with the sampling error for finite-dimensional search spaces V_h, the concept can also be applied for the case of infinite-dimensional search spaces $\mathscr{H}$ without any changes. We will need an $c_\psi \in \mathbb{R}^+$, such that

$$\psi(f(\mathbf{t}),\mathbf{x}) \leq c_\psi \tag{25}$$

for ρ-almost every $(\mathbf{t},\mathbf{x})$ and every $f \in V_{h,b}$. Here, for the case (5), we can e.g. choose

$$c_\psi = (M+r)^2 \tag{26}$$

where $M = \max_{f\in V_{h,b}} \|f\|_{L_{\infty,\rho_T}(T;\mathbb{R}^d)}$ denotes the L_∞ norm bound of functions in $V_{h,b}$. If M is assumed to be bounded independently of h and b (M-boundedness), i.e.

$$\sup_{b>0} \max_{f\in V_{h,b}} \|f\|_{L_{\infty,\rho_T}(T;\mathbb{R}^d)} < \infty, \tag{27}$$

then c_ψ is a universal constant. While (27) does not hold for many choices of V_h, this condition could be enforced by considering $\{f \in V_{h,b} \mid \|f\|_{L_{\infty,\rho_T}(T;\mathbb{R}^d)} \leq \tau\}$ for some constant τ with $\|\hat{f}\|_{L_{\infty,\rho_T}(T;\mathbb{R}^d)} \leq \tau < \infty$ as search set instead of $V_{h,b}$. Without this assumption we achieve

$$c_\psi = (c_{V_h} b + r)^2 \tag{28}$$

as an upper bound for functions in $V_{h,b}$ since $\|f(\mathbf{t})\|_{\ell_2} \leq \|f\|_\infty \leq c_{V_h}\|f\|_{V_h} \leq c_{V_h} b$ and $\|\mathbf{x}\|_{\ell_2} \leq r$ for ρ-almost every $(\mathbf{t},\mathbf{x})$. Note that c_ψ depends quadratically on b without the assumption of M-boundedness.

As mentioned earlier, we usually do not know the measure ρ but are given only a finite sample $\mathscr{Z}_n$ of size n. Therefore, we have to solve (9) over $V_{h,b}$ instead of (2). This gives rise to the so-called *sampling error*

$$\mathscr{E}(f_{\mathscr{Z}_n,V_{h,b}}) - \mathscr{E}(f_{V_{h,b}}), \tag{29}$$

which is just the second part of the error splitting (19).

As we consider functions $f \in V_{h,b} \subset L_{2,\rho_T}(T;\mathbb{R}^d)$ and ψ is continuous on $V_{h,b}$, see (4), we can conclude that $\psi(f(\cdot),\cdot)$ is a ρ-measurable function. Therefore, for an $\eta > 0$, the application of Hoeffding's inequality leads to

$$\mathbb{P}\left[\mathscr{E}(f) - \mathscr{E}_{\mathscr{Z}_n}(f) > \eta\right] \leq \exp\left(-\frac{n\eta^2}{2c_\psi^2}\right) \tag{30}$$

for one particular f. Applying Proposition 3.13 of [9], we obtain[8]

$$\mathbb{P}\left[\sup_{f\in V_{h,b}} |\mathscr{E}(f) - \mathscr{E}_{\mathscr{Z}_n}(f)| > \eta\right] \leq \mathscr{N}\left(V_{h,b}, \frac{\eta}{4C}, L_\infty(T;\mathbb{R}^d)\right) \cdot \exp\left(-\frac{n\eta^2}{8c_\psi^2}\right) \tag{31}$$

[8] A result based on the L_2 covering number instead of L_∞ would be more natural since this reflects the norm in which the overall error is measured, cf. Lemma 1 and Lemma 2. However, it has been shown in [24] that such a result cannot hold in full generality.

where C is the constant from Lemma 1 for $M = \max_{f \in V_{h,b}} \|f\|_{L_{\infty,\rho_T}(T;\mathbb{R}^d)}$ and $\mathcal{N}(X, \varepsilon, Y)$ denotes the covering number of X for balls of radius ε measured in the norm of Y. Using this result, we can now easily derive a bound for the sampling error.

LEMMA 7. *Let C be the constant from Lemma 1 for $M = \max_{f \in V_{h,b}} \|f\|_{L_{\infty,\rho_T}(T;\mathbb{R}^d)}$ and let c_ψ fulfill (25). It holds*

$$\mathbb{P}\left[\mathscr{E}(f_{\mathscr{Z}_n,V_{h,b}}) - \mathscr{E}(f_{V_{h,b}}) > \eta\right] \leq \mathcal{N}\left(V_{h,b}, \frac{\eta}{8C}, L_\infty(T;\mathbb{R}^d)\right) \exp\left(-\frac{n\eta^2}{32c_\psi^2}\right). \tag{32}$$

Proof. The proof works completely analogously to the proof of Lemma 2 in [8], which deals with the special case (5) only. To this end, note that

$$\begin{aligned}
\mathscr{E}(f_{\mathscr{Z}_n,V_{h,b}}) - \mathscr{E}(f_{V_{h,b}}) &= \mathscr{E}(f_{\mathscr{Z}_n,V_{h,b}}) - \mathscr{E}_{\mathscr{Z}_n}(f_{\mathscr{Z}_n,V_{h,b}}) + \mathscr{E}_{\mathscr{Z}_n}(f_{\mathscr{Z}_n,V_{h,b}}) - \mathscr{E}_{\mathscr{Z}_n}(f_{V_{h,b}}) \\
&\quad + \mathscr{E}_{\mathscr{Z}_n}(f_{V_{h,b}}) - \mathscr{E}(f_{V_{h,b}}) \\
&\leq \mathscr{E}(f_{\mathscr{Z}_n,V_{h,b}}) - \mathscr{E}_{\mathscr{Z}_n}(f_{\mathscr{Z}_n,V_{h,b}}) + \mathscr{E}_{\mathscr{Z}_n}(f_{V_{h,b}}) - \mathscr{E}(f_{V_{h,b}}),
\end{aligned}$$

since $f_{\mathscr{Z}_n,V_{h,b}}$ minimizes $\mathscr{E}_{\mathscr{Z}_n}$ over $V_{h,b}$ and therefore $\mathscr{E}_{\mathscr{Z}_n}(f_{\mathscr{Z}_n,V_{h,b}}) - \mathscr{E}_{\mathscr{Z}_n}(f_{V_{h,b}}) \leq 0$. We obtain

$$\begin{aligned}
\mathbb{P}\left[\mathscr{E}(f_{\mathscr{Z}_n,V_{h,b}}) - \mathscr{E}(f_{V_{h,b}}) \leq \eta\right] &\geq \mathbb{P}\left[\mathscr{E}(f_{\mathscr{Z}_n,V_{h,b}}) - \mathscr{E}_{\mathscr{Z}_n}(f_{\mathscr{Z}_n,V_{h,b}}) \leq \frac{\eta}{2}\right. \\
&\qquad \left. \text{and } \mathscr{E}_{\mathscr{Z}_n}(f_{V_{h,b}}) - \mathscr{E}(f_{V_{h,b}}) \leq \frac{\eta}{2}\right] \\
&\geq \mathbb{P}\left[\sup_{f \in V_{h,b}} |\mathscr{E}(f) - \mathscr{E}_{\mathscr{Z}_n}(f)| \leq \frac{\eta}{2}\right] \\
&\overset{(31)}{\geq} 1 - \mathcal{N}\left(V_{h,b}, \frac{\eta}{8C}, L_\infty(T;\mathbb{R}^d)\right) \cdot \exp\left(-\frac{n\eta^2}{32c_\psi^2}\right),
\end{aligned}$$

which proves (32). ∎

If we use the fact that the set $V_{h,b}$ is a convex subset of $C(T;\mathbb{R}^d)$-functions, we can even get rid of the quadratic dependence on η for the cost function (5), i.e.

$$\mathbb{P}\left[\mathscr{E}(f_{\mathscr{Z}_n,V_{h,b}}) - \mathscr{E}(f_{V_{h,b}}) > \eta\right] \leq \mathcal{N}\left(V_{h,b}, \frac{\eta}{12\sqrt{c_\psi}}, L_\infty(T;\mathbb{R}^d)\right) \cdot \exp\left(-\frac{n\eta}{300c_\psi}\right), \tag{33}$$

see Theorem 3.3 of [9]. Moreover, under the premise that the bias is small enough it can be shown that we get rid of the quadratic dependence on η also in the case of non-convex search sets, see Theorem 3.2 of [34].

5. Examples. In this section we will apply the error bounds introduced in the earlier sections to two example settings. Our first example deals with piecewise linear functions with compact support on a sparse grid. The second one considers a spectral method with Fourier polynomials on hyperbolic crosses. Both examples are motivated by the fact that algorithms which employ finite-dimensional grid-based search sets can overcome the problem of (quadratic or even) cubic costs with respect to the amount of data points n which standard data-based approaches inherently suffer from, see chapter 10 of [31]. Therefore, grid-based search sets are a good alternative to kernel methods especially for the case of lower-dimensional problems on large data sets. However, conventional grid based methods encounter the curse of dimensionality, i.e. for an m-dimensional problem the number of degrees of freedom of a tensor product grid is of the order of $\mathcal{O}\left(2^{Lm}\right)$, where 2^L is the resolution

in one direction. To circumvent this issue and to be able to deal with moderate-dimensional (up to $m \approx 20$) problems, the sparse grid ansatz space or the hyperbolic cross space, respectively, is chosen instead of the standard full tensor grid. Regression methods based on these spaces have been successfully applied in the recent years, see e.g. [4],[29].

For both examples we will use the cost function (5), i.e. $\psi(\mathbf{x},\mathbf{y}) := \|\mathbf{x}-\mathbf{y}\|_{\ell_2}^2$. As we deal with functions from $V_{h,b}$ when considering the sampling error, we have the upper bound $c_\psi = (M+r)^2$ with $M = \max_{f\in V_{h,b}} \|f\|_{L_{\infty,\rho_T}(T;\mathbb{R}^d)}$, see (26). Under the condition (27) of M-boundedness, c_ψ is an absolute constant. Without this condition, we have $c_\psi = (c_{V_h} b + r)^2$, see (28). We will also comment on the specific choices of b, see (22), which are needed to apply Theorem 6.

After we fixed our grid-based search spaces V_h, we still have the choice of the regularization norm[9] $\|\cdot\|_{V_h}$. Here, several sparsity-inducing norms of the coefficient vectors (e.g. $\|\cdot\|_{\ell_1}$) and gradient-based norms of the functions (e.g. $\|\cdot\|_{H_1}$) are only a few of the alternatives that can be found in the literature, see e.g. [1],[12]. For our examples, we deliberately choose $\|\cdot\|_{V_h}$ to be the H^1_{mix} Sobolev norm of dominating mixed smoothness which will be introduced in detail in the next subsection. Thus, the embedding constant c_{V_h} can be chosen independently of the discretization parameter h since the norm $\|\cdot\|_{H^1_{\text{mix}}(T;\mathbb{R}^d)}$ is uniformly bounded by $\|\cdot\|_{C(T;\mathbb{R}^d)}$ for any dimension $m \in \mathbb{N}$. This is a main advantage of our choice compared to the standard Sobolev norm $\|\cdot\|_{H^1(T;\mathbb{R}^d)}$ for which this only holds for $m = 1$ due to the Sobolev embedding theorem. Note that the H^1_{mix} norm fits our first example where we deal with sparse grids based on piecewise linear functions[10], see also [4]. Note furthermore that the $H^s_{\text{mix}}(T;\mathbb{R}^d)$ norms which will appear in our examples with $s > 0$ are solely used to measure the smoothness of the solution $\hat{f}$ from (2). They must not be confused with the regularization norm which we have fixed to $\|\cdot\|_{V_h} = \|\cdot\|_{H^1_{\text{mix}}}$.

5.1. Multivariate regression with piecewise linear basis functions on sparse grids. First, let us consider a multivariate regression setting with sparse grids which are based on piecewise linear ansatz functions. To this end, let $T = [0,1]^m$ and let $\rho_T = \lambda_T$ be the Lebesgue measure. For simplicity we will here only deal with the scalar-valued case $d = 1$, the vector-valued case works analogously. We denote by $H^s_{\text{mix}}(T) \subset L_2(T)$ the (Bessel potential) Sobolev space of mixed smoothness of order s, i.e.

$$\|f\|^2_{H^s_{\text{mix}}(T)} := \sum_{\mathbf{i}\in\mathbb{N}^m, \|\mathbf{i}\|_\infty \le s} \|D^{\mathbf{i}} f\|^2_{L_2(T)} \tag{34}$$

for $s \in \mathbb{N}$. For $s \notin \mathbb{N}$ the norm can be defined as restriction of the norm on the whole space $\mathbb{R}^m$ which measures the decay of coefficients of the Fourier transform $\mathscr{F}$, i.e.

$$\|f\|_{H^s_{\text{mix}}(\mathbb{R}^m)} := \|\mathscr{F}^{-1} \prod_{i=1}^m (1+|t_i|^2)^{\frac{s}{2}} \mathscr{F} f\|_{L_2(\mathbb{R}^m)},$$

where t_i denotes the i-th coordinate in $\mathbb{R}^m$. Then

$$\|f\|_{H^s_{\text{mix}}(T)} = \inf_{g\in H^s_{\text{mix}}(\mathbb{R}^m), g|_T = f} \|g\|_{H^s_{\text{mix}}(\mathbb{R}^m)} \tag{35}$$

gives a norm for all $s \in \mathbb{R}^+$ which is equivalent to (34) for $s \in \mathbb{N}$. An alternative is the definition by complex interpolation theory. For details, see [22],[35].

[9] Note at this point that $\|\cdot\|_{V_h}$ can be chosen quite freely since we deal here with regularization and V_h itself is a finite-dimensional space anyway.

[10] For the case of smoother ansatz functions in V_h the choice of Sobolev norms of mixed smoothness of higher degree is also valid.

Discretization error. Now we construct a piecewise linear basis for V_h. Here, we decided to use the so-called piecewise linear prewavelets, see [15], because of their L_2 stability. To this end, let us first define the univariate hat functions

$$\phi(x) := \begin{cases} 1-|x| & \text{if } x \in [-1,1] \\ 0 & \text{else} \end{cases} \qquad \text{and} \quad \phi_{l,i}(x) := 2^{\frac{l}{2}} \phi(2^l \cdot x - i)|_{[0,1]} \tag{36}$$

for $l \in \mathbb{N}$ and $i \in \{0,1,\ldots,2^l-1,2^l\}$. With this definition, we construct the univariate prewavelet basis as follows: Let

$$\gamma_{0,0} := \phi_{0,0},\ \gamma_{0,1} := 2 \cdot \phi_{0,1} - 1,\ \gamma_{1,1} := 2 \cdot \phi_{1,1} - 1.$$

For $l \geq 2$ let $I_l := \left\{ i \in \mathbb{N} \mid 1 \leq i \leq 2^l - 1,\ i \text{ odd} \right\}$ and

$$\gamma_{l,i} := \frac{1}{10}\phi_{l,i-2} - \frac{6}{10}\phi_{l,i-1} + \phi_{l,i} - \frac{6}{10}\phi_{l,i+1} + \frac{1}{10}\phi_{l,i+2}$$

for $i \in I_l, i \neq 1, 2^l - 1$ and

$$\gamma_{l,1} := -\frac{6}{5}\phi_{l,0} + \frac{11}{10}\phi_{l,1} - \frac{3}{5}\phi_{l,2} + \frac{1}{10}\phi_{l,3},\ \ \gamma_{l,2^l-1}(t) := \gamma_{l,1}(1-x).$$

The construction of an m-variate prewavelet function is then straightforward via the tensor product

$$\gamma_{\mathbf{l},\mathbf{i}}(\mathbf{t}) := \prod_{j=1}^{m} \gamma_{j,i_j}(t_j),$$

where $\mathbf{l} = (l_1,\ldots,l_m) \in \mathbb{N}^m$ is the multivariate level and $\mathbf{i} = (i_1,\ldots,i_m) \in \mathbb{N}^m$ denotes the multivariate position index. Now let

$$\mathbf{I}_\mathbf{l} := \left\{ \mathbf{i} \in \mathbb{N}^m \;\middle|\; \begin{array}{ll} 0 \leq i_j \leq 1, & \text{if } l_j = 0 \\ 1 \leq i_j \leq 2^{l_j} - 1,\ i_j \text{ odd} & \text{if } l_j > 0 \end{array} \quad \text{for all } 1 \leq j \leq m \right\}. \tag{37}$$

Then, $W_\mathbf{l} := \operatorname{span}\left\{ \gamma_{\mathbf{l},\mathbf{i}} \mid \mathbf{i} \in \mathbf{I}_\mathbf{l} \right\}$ is the so-called hierarchical increment space (or detail space) of level $\mathbf{l}$. We define

$$V_\mathbf{l} := \bigoplus_{\mathbf{k} \leq \mathbf{l}} W_\mathbf{k} = \operatorname{span}\{B_\mathbf{l}\} \tag{38}$$

with the hierarchical prewavelet basis

$$B_\mathbf{l} := \left\{ \gamma_{\mathbf{k},\mathbf{i}} \mid \mathbf{i} \in \mathbf{I}_\mathbf{k}, \mathbf{k} \leq \mathbf{l} \right\},$$

where $\mathbf{k} \leq \mathbf{l}$ has to be understood elementwise. For more details on the construction and the properties of multivariate prewavelet bases, see [15].

Now we are able to define the sparse grid space of level $L > 0$ by

$$V_h = V^L := \bigoplus_{\substack{\mathbf{k} \in \mathbb{N}^m \\ \zeta_m(\mathbf{k}) \leq L}} W_\mathbf{k}\ , \tag{39}$$

where $h = 2^{-L}$ denotes the minimal mesh width. Here, $\zeta_m(\mathbf{0}) := 0$ and

$$\zeta_m(\mathbf{k}) := |\mathbf{k}|_1 - m + \left|\{ j \mid \mathbf{k}_j = 0 \}\right| + 1$$

for every other $\mathbf{k} \in \mathbb{N}^m$. This specific definition of ζ_m guarantees that the resolution of grids on the boundary is the same as the resolution of grids in the interior of the domain. The dimension of V_h is bounded from above by

$$\dim(V_h) = 3^m \cdot \left(2^L \frac{L^{m-1}}{(m-1)!} + \mathscr{O}\left(L^{m-2}\right) \right) = \mathscr{O}\left(2^L L^{m-1}\right), \tag{40}$$

see e.g. [11]. The sparse grid space (39) copes with the curse of dimensionality which is induced by the dimension m. Note here that a properly adjusted sparse grid space might be more appropriate if the regression function is known to belong to a mixed smoothness class of degree larger than 2 or to a mixed smoothness class of varying degrees for different directions, see [13]. Furthermore, adaptive sparse grids can be employed to deal with non-smooth solutions. An exhaustive consideration of appropriate sparse grid discretization spaces V_h is beyond the scope of this paper, but see e.g. [5],[22] for details in this direction.

In [4] the regularization was realized in the Sobolev space of mixed smoothness $H^1_{\text{mix}}(T)$. Note again that then the embedding constant c_{V_h} is bounded from above independently of the discretization parameter h for arbitrary dimension m.

For the prewavelet basis it is known, see [15], that there exists a $c_m > 0$ depending only on m such that

$$\|f\|_{V_h} \leq c_m 2^L \|f\|_{L_2(T)}$$

for all $f \in V_h$. Therefore, the Bernstein factor from Lemma 5 can be chosen as $c(h) = c_m h^{-1} = c_m 2^L$.

We now apply Lemma 5 and get

$$\|f_h^{\text{BA}}\|_{V_h} \leq c_m 2^L \|f_\rho\|_{L_2(T)} =: b \tag{41}$$

for the best L_2 approximation f_h^{BA} to f_ρ from V_h. Thus, from Lemma 5 and Theorem 6, we obtain

$$\mathscr{E}(f_{V_{h,b}}) - \mathscr{E}(f_\rho) = \inf_{f \in V_h} \|f - f_\rho\|^2_{L_2(T)}.$$

The L_2 error rate for the approximation of $f_\rho \in H^s_{\text{mix}}(T)$ for some $0 \leq s \leq 2$ with piecewise linear splines on a sparse grid[11] of level L is known to be bounded by

$$\inf_{f \in V_h} \|f - f_\rho\|_{L_2(T)} \leq \mathscr{O}\left(2^{-sL} L^{m-1}\right),$$

for $L \to \infty$, see Theorem 4 of [22]. Note that only the periodic case has been treated there. However, to prove the result for the non-periodic case just the Riesz-stability of the basis and specific Jackson and Bernstein inequalities are needed, all of which are fulfilled in our case, see [5],[15],[16]. Thus, Theorem 4 of [22] also holds.

Sampling error. We employ the bound (33) with functions from the ball $V_{h,b} \subset V_h$, i.e.

$$P := \mathbb{P}\left[\mathscr{E}(f_{\mathscr{Z}_n, V_{h,b}}) - \mathscr{E}(f_{V_{h,b}}) > \eta \right] \leq \mathscr{N}\left(V_{h,b}, \frac{\eta}{12\sqrt{c_\psi}}, L_\infty(T; \mathbb{R}^d) \right) \exp\left(-\frac{n\eta}{300 c_\psi} \right)$$

with c_ψ from (26). First, we have to provide an upper bound for the covering number. Let us denote the degrees of freedom in V_h by $N := \dim(V_h)$.

[11] Note that the spline space spanned by the piecewise linear hat functions on a sparse grid of level L and the prewavelet space V_h coincide.

LEMMA 8. *For $b > 0$ bounded away from zero and $\eta > 0$ bounded from above there exists a constant $c_{\mathscr{N}}$ such that*

$$\mathscr{N}\left(V_{h,b}, \frac{\eta}{12\sqrt{c_\psi}}, L_\infty(T;\mathbb{R})\right) \leq \left(\frac{c_{\mathscr{N}} c_{V_h} \sqrt{c_\psi} b}{\eta}\right)^N.$$

Proof. We have $\|\cdot\|_{L_\infty} \leq c_{V_h} \|\cdot\|_{V_h}$ for functions in V_h. Therefore, every $\frac{\varepsilon}{c_{V_h}}$ covering[12] with respect to the V_h norm is an ε covering with respect to the $L_\infty(T;\mathbb{R})$ norm and we have

$$\begin{aligned}\mathscr{N}\left(V_{h,b}, \frac{\eta}{12\sqrt{c_\psi}}, L_\infty(T;\mathbb{R})\right) &\leq \mathscr{N}\left(V_{h,b}, \frac{\eta}{12 c_{V_h}\sqrt{c_\psi}}, V_h\right) \\ &\leq \left(\frac{24 c_{V_h}\sqrt{c_\psi} b}{\eta} + 1\right)^N.\end{aligned}$$

The last inequality[13] is proven in Theorem 5.3 from [9]. It holds for any finite-dimensional Banach space V_h. As $\frac{b}{\eta}$ is bounded away from zero there exists a $c_{\mathscr{N}}$ such that

$$\left(\frac{24 c_{V_h}\sqrt{c_\psi} b}{\eta} + 1\right)^N \leq \left(\frac{c_{\mathscr{N}} c_{V_h}\sqrt{c_\psi} b}{\eta}\right)^N. \qquad \square$$

Note that the prerequisites of Lemma 8 are no restriction for our analysis as we are interested in the case $b \to \infty$ and $\eta \to 0$. Applying the Lemma we have

$$\begin{aligned} P &\leq \left(c_{\mathscr{N}} c_{V_h}\sqrt{c_\psi}\frac{b}{\eta}\right)^N \exp\left(-\frac{n\eta}{300 c_\psi}\right) \\ \Leftrightarrow\ \exp\left(\frac{n\eta}{300 c_\psi}\right)\eta^N &\leq \left(c_{\mathscr{N}} c_{V_h}\sqrt{c_\psi} b P^{-\frac{1}{N}}\right)^N \\ \Leftrightarrow\ \exp(\alpha\eta)\,\eta &\leq \beta \end{aligned} \tag{42}$$

with $\alpha := \frac{n}{300 c_\psi N}$ and $\beta := c_{\mathscr{N}} c_{V_h}\sqrt{c_\psi} b P^{-\frac{1}{N}}$. Therefore, applying the monotone Lambert function $W : [0,\infty) \to [0,\infty)$ defined by

$$W(t\exp(t)) = t$$

on both sides of (42), we obtain

$$\begin{aligned} &\alpha\eta \leq W(\alpha\beta) \leq \max(1, \log(\alpha\beta)) \\ \Leftrightarrow\ &\eta \leq \frac{1}{\alpha}\max(1, \log(\alpha\beta)). \end{aligned}$$

The probability P can be interpreted as a function $P(\eta, N, b, n)$. Let us now choose $0 < \delta < 1$ such that the following assumption[14] holds: There exists an $\varepsilon > 0$ such that jumps of the

[12] A subset $B \subset A$ of points such that for every $a \in A$ there exists a $b \in B$ which fulfills $\|a - b\| \leq \delta$ is called a δ covering of the set A with respect to the norm $\|\cdot\|$.

[13] Note that our estimate for the covering number is not exploiting the fact that the norm $\|\cdot\|_{V_h}$ is stronger than $\|\cdot\|_{L_\infty}$.

[14] Note that apart from the trivial case where the sampling error is already 0 for most choices of $\mathscr{Z}_n$, the assumption of small (or zero) jump sizes for large enough N, b and n is quite natural since $\mathscr{E}$ is continuous and the measure $\rho(T \times \cdot)$ is usually not singular.

function P along direction η are of size $\delta-\varepsilon$ or less for large enough N,b and n, i.e. there exist $\varepsilon,N_0,b_0,n_0>0$ such that

$$\lim_{\eta\nearrow\bar{\eta}} P(\eta,N,b,n) - \lim_{\eta\searrow\bar{\eta}} P(\eta,N,b,n) \leq \delta-\varepsilon \quad \forall\, \bar{\eta}>0, N>N_0, b>b_0, n>n_0. \tag{43}$$

Now let us choose η_δ as the smallest η (depending on N,b,n) such that $P(\eta,N,b,n)\leq\delta$. Because of assumption (43) we know that

$$\delta-\varepsilon \leq P(\eta_\delta,N,b,n)\leq\delta$$

and therefore $P(\eta_\delta,N,b,n)^{-1/N}\leq(\delta-\varepsilon)^{-1/N}$ is bounded from above for all $N\geq 1$. Altogether we obtain the result that with probability at least $1-\delta$ we have

$$\mathscr{E}(f_{\mathscr{Z}_n,V_{h,b}})-\mathscr{E}(f_{V_{h,b}})\leq\eta_\delta=\mathscr{O}\left(c_\psi\frac{N}{n}\max\left(\log\left(\frac{nbc_{V_h}}{\sqrt{c_\psi}N}\right),1\right)\right), \tag{44}$$

with the factor $-\log(\delta-\varepsilon)$ appearing in the $\mathscr{O}$-constant.

On the assumption that we have M-boundedness, it holds $c_\psi=\mathscr{O}(1)$, see (27). Therefore, we obtain

$$\eta_\delta=\mathscr{O}\left(\frac{N}{n}\max\left(\log\left(\frac{bn}{N}\right),1\right)\right).$$

Substituting $b\sim 2^L$ and $N\sim 2^L L^{m-1}$, see (41) and (40), we get

$$\eta_\delta=\mathscr{O}\left(\frac{2^L L^{m-1}}{n}\max\left(\log\left(\frac{n}{L^{m-1}}\right),1\right)\right). \tag{45}$$

If we use (28) instead we choose the embedding constant c_{V_h} independently of the discretization since $\|\cdot\|_{V_h}=\|\cdot\|_{H^1_{\mathrm{mix}}(T)}$. This leads to $c_\psi=\mathscr{O}(b^2)$ and $\frac{b}{\sqrt{c_\psi}}=\mathscr{O}(1)$ for $b\to\infty$. Thus, (44) becomes

$$\eta_\delta=\mathscr{O}\left(b^2\frac{N}{n}\max\left(\log\left(\frac{n}{N}\right),1\right)\right).$$

Finally, substituting $b\sim 2^L$ and $N\sim 2^L L^{m-1}$, we get

$$\eta_\delta=\mathscr{O}\left(\frac{2^{3L}L^{m-1}}{n}\max\left(\log\left(\frac{n}{2^L L^{m-1}}\right),1\right)\right) \tag{46}$$

in L and n.

Overall rate. Now we choose the largest $0<\hat{s}\leq 2$ such that $f_\rho\in H^{\hat{s}}_{\mathrm{mix}}(T)\subset L_2(T)$. Furthermore, let $\|\cdot\|_{V_h}$ be the $H^1_{\mathrm{mix}}(T)$ norm and let $b=c_m 2^L\|f_\rho\|_{L_2}$ with a specific constant c_m depending only on m. Then, using the results from the previous subsections, we add discretization error and sampling error to obtain the rate

$$\mathscr{E}(f_{\mathscr{Z}_n,V_{h,b}})-\mathscr{E}(\hat{f})=\mathscr{O}\left(2^{-2\hat{s}L}L^{2(m-1)}+\frac{2^L L^{m-1}}{n}\max\left(\log\left(\frac{n}{L^{m-1}}\right),1\right)\right) \tag{47}$$

under the assumption of M-boundedness, see (27), and the rate

$$\mathscr{E}(f_{\mathscr{Z}_n,V_{h,b}})-\mathscr{E}(\hat{f})=\mathscr{O}\left(2^{-2\hat{s}L}L^{2(m-1)}+\frac{2^{3L}L^{m-1}}{n}\max\left(\log\left(\frac{n}{2^L L^{m-1}}\right),1\right)\right) \tag{48}$$

without M-boundedness. These rates for the overall error are obtained for $L, n \to \infty$ by fixing a confidence $1-\delta$ and using the assumption (43) which leads to a factor $-\log(\delta-\varepsilon)$ in the $\mathcal{O}$-constant. Here, the first term $2^{-2\hat{s}L}L^{2(m-1)}$ of (48) resembles the best approximation error in $H^{\hat{s}}_{\mathrm{mix}}(T)$.

Note at this point that another way to estimate the discretization error from above is to apply Theorem 4.1 of [32]. It gives a rate for the discretization error for any finite-dimensional space V_h under the assumption that certain Jackson and Bernstein inequalities hold. In the case of sparse grids, these inequalities can be found in e.g. [22]. Then, Theorem 4.1 of [32] states that the discretization error can asymptotically be bounded by $2^{-2\hat{s}L+\varepsilon}$ for any $\varepsilon > 0$. This comes close to our result. However, the theorem requires the discretization to be conforming, i.e. $V_h \subset H^{\hat{s}}_{\mathrm{mix}}(T)$ and can not be applied for $\frac{3}{2} \leq \hat{s} \leq 2$ since $V_h \not\subseteq H^{\hat{s}}_{\mathrm{mix}}(T)$ for $\hat{s} \geq \frac{3}{2}$.

Now, let us consider (47) in more detail. The first term of (47) only depends on the discretization parameter L and not on the number of data n. It thus converges to 0 with $L \to \infty$ for any value of n. But the second term depends on both L and n and may diverge. In the limit case $L, n \to \infty$ it can only go to zero if the maximum in the second term becomes $\log\left(\frac{n}{L^{m-1}}\right)$. Then, we have to fulfill the additional, necessary and sufficient condition

$$2^L L^{m-1} \log\left(\frac{n}{L^{m-1}}\right) \stackrel{!}{=} o(n) \tag{49}$$

to achieve convergence of the second term for $L, n \to \infty$. Note that similar conditions are derived in [6],[7],[18],[27] for other examples. Most of these results also show that n has to grow faster (up to a log-factor) than the basis size (in our case $2^L L^{(m-1)}$) to achieve a stable and convergent method.

An analogous consideration for (48) leads to the necessary and sufficient condition

$$2^{3L} L^{m-1} \log\left(\frac{n}{2^L L^{m-1}}\right) \stackrel{!}{=} o(n) \tag{50}$$

to achieve convergence of the second term of (48). Since we did not use M-boundedness here, we cannot rely on fixed L_∞ bounds of the approximant and the scaling of (28) influences the error bound, see (44). Therefore, the condition (50) is more restrictive than (49).

Balancing the overall error. To balance the overall error we choose L and n such that the two summands are approximately of equal size. We here omit the computational costs when balancing the error terms since the mathematical derivation of the optimal choice of n and L for the corresponding cost-benefit-ratio is quite involved and beyond the scope of this paper. Let us first consider the rate (47) under the assumption of M-boundedness, see (27). We will stick to the convergent case and therefore have to couple L and n such that

$$2^{-2\hat{s}L} L^{2(m-1)} \approx \frac{2^L L^{m-1}}{n} \log\left(\frac{n}{L^{m-1}}\right),$$

which leads to

$$n \approx 2^{(1+2\hat{s})L} L^{-m+1} (\log(n) - (m-1)\log(L)).$$

Thus, up to logarithmic factors in the basis size and the amount of sample points, the optimal scaling is

$$n \approx 2^{(1+2\hat{s})L}.$$

Substituting this into (47) we obtain an overall rate

$$\mathscr{E}(f_{\mathscr{X}_n, V_{h,b}}) - \mathscr{E}(\hat{f}) \approx \mathcal{O}\left(n^{-\frac{2\hat{s}}{1+2\hat{s}}}\right)$$

up to logarithmic factors. Thus, for the case of maximal smoothness, i.e. $\hat{s} = 2$, the resulting rate is $\mathscr{O}\left(n^{-\frac{4}{5}}\right)$. Note that this corresponds to the best possible rate which can be achieved at all since already for the simple case of univariate linear spline functions this rate is observed, see e.g. [7, 18]. Moreover, note that, in our m-dimensional case, the exponential dependence on m only affects the logarithmic terms (as it is common for sparse grids). This is in contrast to e.g. [18] where the overall rate deteriorates to $n^{-\frac{4}{m+4}}$ for non-regularized regression on a full grid of multivariate linear tensor product splines.

Without the assumption of M-boundedness, we have to balance

$$2^{-2\hat{s}L} L^{2(m-1)} \approx \frac{2^{3L} L^{m-1}}{n} \log\left(\frac{n}{2^L L^{m-1}}\right),$$

and get

$$n \approx 2^{(3+2\hat{s})L} L^{-m+1} (\log(n) - L\log(2) - (m-1)\log(L)).$$

Here the optimal scaling is

$$n \approx 2^{(3+2\hat{s})L}.$$

up to logarithms. Therefore, the sample size n has to grow 2^{2L} times faster than under the assumption of M-boundedness. Substituting this into (48) leads to

$$\mathscr{E}(f_{\mathscr{Z}_n, V_{h,b}}) - \mathscr{E}(\hat{f}) \approx \mathscr{O}\left(n^{-\frac{2\hat{s}}{3+2\hat{s}}}\right)$$

up to logarithmic factors. Therefore, the smooth case $\hat{s} = 2$ results in the rate $\mathscr{O}\left(n^{-\frac{4}{7}}\right)$.

5.2. Multivariate periodic regression for Fourier polynomials on hyperbolic crosses. In the last example we dealt with the so-called h-version of sparse grids where the degree of the (piecewise) polynomials was fixed and their support was refined. In the following, we consider global Fourier-polynomials where we increase the maximum frequency. This corresponds to a spectral/p-version. Here, we consider a multivariate regression setting with polynomials on hyperbolic crosses. To this end, let $T = [-\pi, \pi]^m$ where we identify opposite hyperplanes and let $\rho_T = \frac{\lambda_T}{(2\pi)^m}$ be the rescaled Lebesgue measure. Furthermore, we assume that f_ρ is 2π-periodic in every coordinate. Again, we deal with the scalar-valued case $d = 1$. We now write $\bar{H}^s_{\text{mix},\rho_T}(T)$ to denote the periodic Sobolev space on T for $0 < s < \infty$. Similar to (35), the norm is defined by

$$\|f\|_{\bar{H}^s_{\text{mix},\rho_T}(T)} := \left\| \sum_{\mathbf{k}\in\mathbb{Z}^m} c_{\mathbf{k}}(f) \prod_{j=1}^m (1 + |k_j|^2)^{\frac{s}{2}} e^{i\mathbf{k}^T\mathbf{t}} \right\|_{L_{2,\rho_T}(T)},$$

where $\mathbf{t} \in T$ is the spatial variable and

$$c_{\mathbf{k}}(f) := \frac{1}{(2\pi)^m} \int_T f(\mathbf{t}) e^{-i\mathbf{k}^T\mathbf{t}} \mathrm{d}\mathbf{t}$$

denotes the $\mathbf{k}$-th Fourier coefficient.

Discretization error. Following [33], let

$$V_h = V_{2^{-L}} := T(\Gamma_L) := \left\{ f \in L_{2,\rho_T}(T) \,\middle|\, f(\mathbf{t}) = \sum_{\mathbf{k}\in\Gamma_L} \alpha_{\mathbf{k}} \exp\left(i\mathbf{k}^T\mathbf{t}\right) \right\}, \tag{51}$$

where

$$\Gamma_L := \left\{ \mathbf{k} \in \mathbb{Z}^m \,\middle|\, \sum_{j=1}^{m} \log_2(\max(|\mathbf{k}_j|, 1)) \leq L \right\}$$

is the hyperbolic cross of level L. Similarly to the sparse grid construction in the last example, $N := \dim(V_h)$ is bounded from above by

$$\dim(V_h) = \mathcal{O}\left(2^L L^{m-1}\right),$$

see [33]. Analogously to the first example, we choose $\|\cdot\|_{V_h} = \|\cdot\|_{\bar{H}^1_{\text{mix},\rho_T}(T)}$ as regularization norm. For the hyperbolic cross construction above, it thus holds

$$\|f\|_{V_h} = \|f\|_{\bar{H}^1_{\text{mix},\rho_T}(T)} \leq c_m 2^L \|f\|_{L_{2,\rho_T}(T)}$$

for all $f \in V_h$, see Theorem III.2.3 of [33]. Therefore, $c(h) = c_m h^{-1} = c_m 2^L$ is a valid choice in Lemma 5. Applying Theorem 6 for the corresponding $b := c(h)\|f_\rho\|_{L_{2,\rho_T}(T)}$ we obtain

$$\mathcal{E}(f_{V_{h,b}}) - \mathcal{E}(f_\rho) = \inf_{f \in V_h} \|f - f_\rho\|^2_{L_{2,\rho_T}(T)}.$$

The rate of the L_2 best approximation error for $f_\rho \in \bar{H}^s_{\text{mix},\rho_T}(T)$ with $0 < s < \infty$ by functions in the hyperbolic cross space is bounded by

$$\inf_{f \in V_h} \|f - f_\rho\|_{L_{2,\rho_T}(T)} = \mathcal{O}\left(2^{-sL}\right),$$

for $L \to \infty$, see Theorem III-3.2 of [33].[15]

Sampling error. The sampling error bound is derived analogously to the last example. Assuming that (43) holds we obtain

$$\eta_\delta = \mathcal{O}\left(\frac{2^L L^{m-1}}{n} \max\left(\log\left(\frac{n}{L^{m-1}}\right), 1\right)\right)$$

under the assumption of M-boundedness from (45) and

$$\eta_\delta = \mathcal{O}\left(\frac{2^{3L} L^{m-1}}{n} \max\left(\log\left(\frac{n}{2^L L^{m-1}}\right), 1\right)\right)$$

without the assumption of M-boundedness from (46), respectively.

Overall rate. Let $0 < \hat{s} < \infty$ be the largest real number such that $f_\rho \in \bar{H}^{\hat{s}}_{\text{mix},\rho_T}(T)$. We choose $\|\cdot\|_{V_h} = \|\cdot\|_{\bar{H}^1_{\text{mix},\rho_T}(T)}$ and $b = c_m 2^L \|f_\rho\|_{L_{2,\rho_T}}$, with a constant c_m which depends only on m. Then, fixing a confidence $1 - \delta$ which fulfills (43), we add discretization error and sampling error to obtain the rate

$$\mathcal{E}(f_{\mathcal{Z}_n,V_{h,b}}) - \mathcal{E}(\hat{f}) = \mathcal{O}\left(2^{-2\hat{s}L} + \frac{2^L L^{(m-1)}}{n} \max\left(\log\left(\frac{n}{L^{m-1}}\right), 1\right)\right) \tag{52}$$

[15] The function classes $\mathbf{MW}^s_2$ in [33] correspond to the unit ball of our spaces $\bar{H}^s_{\text{mix}}(T)$, see section 2.7 of [36] for a thorough proof.

under the assumption of M-boundedness, see (27), and

$$\mathscr{E}(f_{\mathscr{Z}_n,V_{h,b}}) - \mathscr{E}(\hat{f}) = \mathscr{O}\left(2^{-2\hat{s}L} + \frac{2^{3L}L^{(m-1)}}{n}\max\left(\log\left(\frac{n}{2^L L^{m-1}}\right),1\right)\right) \tag{53}$$

without the assumption of M-boundedness. Again, a factor of $-\log(\delta-\varepsilon)$ enters the $\mathscr{O}$-constants. Note here that, since $V_h \subset \bar{H}^{\hat{s}}_{\mathrm{mix},\rho_T}(T)$ for any $\hat{s}>0$, a similar result can be shown by applying Theorem 4.1 of [32].

Analogously to the discussion in the previous example, we will now consider (52) in more detail. Again, the first term of (52) only depends on L and not on n. It thus converges to 0 with $L\to\infty$ for any value of n. The second term may however diverge since it depends on both L and n. Let us now consider the limit case $L,n\to\infty$. We can use the same arguments as in the last example. In the convergent case, i.e. if (52) approaches 0, the maximum term in (52) becomes $\log\left(\frac{n}{L^{m-1}}\right)$. Therefore, a necessary and sufficient condition on the amount of samples n to achieve convergence is

$$2^L L^{m-1}\log\left(\frac{n}{L^{m-1}}\right) \overset{!}{=} o(n)$$

for $L,n\to\infty$.

The same consideration for (53) leads to

$$2^{3L}L^{m-1}\log\left(\frac{n}{2^L L^{m-1}}\right) \overset{!}{=} o(n)$$

for $L,n\to\infty$. without the assumption of M-boundedness.

Balancing the overall error. Analogously to the last example we will now balance the overall error in the convergent case. Under the assumption of M-boundedness a comparison of the discretization error and the sampling error yields

$$2^{-2\hat{s}L} \approx \frac{2^L L^{m-1}}{n}\log\left(\frac{n}{L^{m-1}}\right)$$

which leads to

$$n \approx 2^{(1+2\hat{s})L}L^{m-1}(\log(n)-(m-1)\log(L)).$$

Therefore, the optimal scaling is

$$n \approx 2^{(1+2\hat{s})L}$$

up to logarithms in the basis size $N=\mathscr{O}(2^L L^{m-1})$ or the sample size n. Substituting this into (52) we again obtain

$$\mathscr{E}(f_{\mathscr{Z}_n,V_{h,b}}) - \mathscr{E}(\hat{f}) \approx \mathscr{O}\left(n^{-\frac{2\hat{s}}{1+2\hat{s}}}\right).$$

Thus, the main rate is independent of m and the curse of dimensionality only appears in the logarithmic terms.

Without the assumption of M-boundedness we obtain the scaling

$$n \approx 2^{(3+2\hat{s})L}$$

and with this the rate

$$\mathscr{E}(f_{\mathscr{Z}_n,V_{h,b}}) - \mathscr{E}(\hat{f}) \approx \mathscr{O}\left(n^{-\frac{2\hat{s}}{3+2\hat{s}}}\right).$$

6. Concluding remarks. In this article we introduced the discretization error for regression problems with finite-dimensional search space V_h. We summarized recent developments on the bias and the sampling error and explained why these results cannot directly be applied to the finite-dimensional setting. We coupled the norm bound of functions from the search set $V_{h,b}$ to the discretization level by exploiting a Bernstein inequality in V_h. The corresponding sampling error had to be estimated for $V_{h,b}$. Finally, two examples for multivariate regression concluded the article. For both, piecewise linear splines on sparse grids and Fourier polynomials on hyperbolic crosses, we have shown that an optimal scaling in the sample size n and the dimension of the finite-dimensional ansatz space N leads, up to logarithmic terms, to the same convergence rate which is obtained for the univariate case.

An actual algorithm usually employs the method of Lagrangian multipliers instead of dealing with a constrained optimization problem. This alternative has not been discussed here. However, there is a direct relation between the primal (constrained optimization) problem and the dual (Lagrangian multiplier) problem, see e.g. [9]. Furthermore, more general methods can be considered as well, e.g. Hilbert scales for Tikhonov regularization, see [20].

Modifications of the methods introduced in this paper can be used to treat the case of principal manifold learning, see [19],[31].

Finally, the derivation of tight upper bounds for specific settings as well as generic lower bounds for the overall error in regression will be a task for future research. To this end, note that e.g. in the setting of noise-free regression without explicit function norm regularization, our results can be improved for specific finite-dimensional function classes, see e.g. [6],[7],[27].

REFERENCES

[1] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, *Optimization with Sparsity-Inducing Penalties*, Now Publishers Inc., 2011.
[2] P. Binev, A. Cohen, W. Dahmen, R. DeVore, and V. Temlyakov, *Universal algorithms for learning theory - part I: piecewise constant functions*, Journal of Machine Learning Research, 6 (2005), pp. 1297–1321.
[3] P. Binev, A. Cohen, W. Dahmen, R. DeVore, and V. Temlyakov, *Universal algorithms for learning theory - part II: piecewise polynomial functions*, Constructive Approximation, 26 (2007), pp. 127–152.
[4] B. Bohn and M. Griebel, *An adaptive sparse grid approach for time series predictions*, in Sparse grids and applications, J. Garcke and M. Griebel, eds., vol. 88 of Lecture Notes in Computational Science and Engineering, Springer, 2012, pp. 1–30.
[5] H.-J. Bungartz and M. Griebel, *Sparse grids*, Acta Numerica, 13 (2004), pp. 147–269.
[6] A. Chkifa, A. Cohen, G. Migliorati, F. Nobile, and R. Tempone, *Discrete least squares polynomial approximation with random evaluations - application to parametric and stochastic elliptic PDEs*, ESAIM: Mathematical Modelling and Numerical Analysis (M2AN), 49 (2015), pp. 815–837.
[7] A. Cohen, M. Davenport, and D. Leviatan, *On the stability and accuracy of least squares approximations*, Foundations of Computational Mathematics, 13 (2013), pp. 819–834.
[8] F. Cucker and S. Smale, *On the mathematical foundations of learning*, Bulletin of the American Mathematical Society, 39 (2001), pp. 1–49.
[9] F. Cucker and D. Zhou, *Learning Theory*, Cambridge Monographs on Applied and Computational Mathematics, 2007.
[10] R. DeVore and G. Lorentz, *Constructive Approximation*, A Series of Comprehensive Studies in Mathematics, Springer, 1993.
[11] C. Feuersänger, *Sparse Grid Methods for Higher Dimensional Approximation*, PhD thesis, Institute for Numerical Simulation, University of Bonn, 2010.
[12] J. Garcke and M. Hegland, *Fitting multidimensional data using gradient penalties and the sparse grid combination technique*, Computing, 84 (2009), pp. 1–25.
[13] M. Griebel and H. Harbrecht, *On the construction of sparse tensor product spaces*, Mathematics of Computations, 82 (2013), pp. 975–994.
[14] M. Griebel and M. Hegland, *A finite element method for density estimation with Gaussian priors*, SIAM Journal on Numerical Analysis, 47 (2010), pp. 4759–4792.
[15] M. Griebel and P. Oswald, *Tensor product type subspace splitting and multilevel iterative methods for*

anisotropic problems, Adv. Comput. Math., 4 (1995), pp. 171–206.
[16] M. GRIEBEL, P. OSWALD, AND T. SCHIEKOFER, *Sparse grids for boundary integral equations*, Numerische Mathematik, 83 (1999), pp. 279–312.
[17] M. GRIEBEL, C. RIEGER, AND B. ZWICKNAGL, *Multiscale approximation and reproducing kernel Hilbert space methods*, SIAM Journal on Numerical Analysis, 53 (2015), pp. 852–873.
[18] L. GYÖRFI, M. KOHLER, A. KRZYŻAK, AND H. WALK, *A Distribution-Free Theory of Nonparametric Regression*, Springer, 2002.
[19] T. HASTIE AND W. STUETZLE, *Principal curves*, Journal of the American Statistical Association, 84 (1989), pp. 502–516.
[20] M. HEGLAND, *An optimal order regularization method which does not use additional smoothness assumptions*, SIAM J. Numer. Anal., 29 (1992), pp. 1446–1461.
[21] M. HEGLAND, *Data mining techniques*, Acta Numerica, 10 (2001), pp. 313–355.
[22] S. KNAPEK, *Approximation und Kompression mit Tensorprodukt-Multiskalenräumen*, PhD thesis, University of Bonn, 2000.
[23] M. KOHLER, *Inequalities for uniform deviations of averages from expectations with applications to nonparametric regression*, Journal of Statistical Planning and Inference, 89 (2000), pp. 1–23.
[24] S. KONYAGIN AND V. TEMLYAKOV, *The entropy in learning theory. Error estimates*, Constructive Approximation, 25 (2007), pp. 1–27.
[25] J. A. LEE AND M. VERLEYSEN, *Nonlinear Dimensionality Reduction*, Springer Science, 2007.
[26] C. MICCHELLI AND M. PONTIL, *On learning vector-valued functions*, Neural Computation, 17 (2005), pp. 177–204.
[27] G. MIGLIORATI, F. NOBILE, E. VON SCHWERIN, AND R. TEMPONE, *Analysis of discrete L^2 projection on polynomial spaces with random evaluations*, Foundations of Computational Mathematics, 14 (2014), pp. 419–456.
[28] J. PEETRE, *A theory of interpolation of normed spaces*, vol. 39 of Notas de Matemática, Rio de Janeiro: Instituto de Matemática Pura e Aplicada, Conselho Nacional de Pesquisas, 1968.
[29] D. PFLÜGER, B. PEHERSTORFER, AND H.-J. BUNGARTZ, *Spatially adaptive sparse grids for high-dimensional data-driven problems*, Journal of Complexity, 26 (2010), pp. 508–522.
[30] T. POGGIO, R. RIFKIN, S. MUKHERJEE, AND P. NIYOGI, *General conditions for predictivity in learning theory*, Nature, 428 (2004), pp. 419–422.
[31] B. SCHÖLKOPF AND A. SMOLA, *Learning with Kernels – Support Vector Machines, Regularization, Optimization, and Beyond*, The MIT Press – Cambridge, Massachusetts, 2002.
[32] S. SMALE AND D. ZHOU, *Estimating the approximation error in learning theory*, Analysis and Applications, 1 (2003), pp. 17–41.
[33] V. TEMLYAKOV, *Approximation of Periodic Functions*, Nova Science, 1993.
[34] V. TEMLYAKOV, *Approximation in learning theory*, Constructive Approximation, 27 (2008), pp. 33–74.
[35] H. TRIEBEL, *Interpolation Theory, Function Spaces, Differential Operators*, North Holland Mathematical Library, 1978.
[36] T. ULLRICH, *Smolyaks Algorithm, Sparse Grid Approximation and Periodic Function Spaces with Dominating Mixed Smoothness*, PhD thesis, University of Jena, 2007.
[37] V. VAPNIK, *Statistical Learning Theory*, John Wiley & Sons, 1998.
[38] M. WONG AND M. HEGLAND, *Maximum a posteriori density estimation and the sparse grid combination technique*, in Proceedings of the 16th Biennial Computational Techniques and Applications Conference, CTAC-2012, S. McCue, T. Moroney, D. Mallet, and J. Bunder, eds., vol. 54, 2013, pp. 508–522.
[39] Q. WU AND D. ZHOU, *Learning with sample dependent hypothesis spaces*, Computers & Mathematics with Applications, 56 (2008), pp. 2896–2907.
[40] Y. YING AND D. ZHOU, *Learnability of Gaussians with flexible variances*, Journal of Machine Learning Research, 8 (2007), pp. 249–276.
[41] Y. ZHANG, F. CAO, AND Z. XU, *Estimation of learning rate of least square algorithm via Jackson operator*, Neurocomputing, 74 (2011), pp. 516–521.