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

On the convergence rate of sparse grid least squares regression

INS Preprint No. 1711

August 2017

On the convergence rate of sparse grid least squares regression

Bastian Bohn

Abstract While sparse grid least squares regression algorithms have been frequently used to tackle Big Data problems with a huge number of input data in the last 15 years, a thorough theoretical analysis of stability properties, error decay behavior and appropriate couplings between the dataset size and the grid size has not been provided yet.

In this paper, we will present a framework which will allow us to close this gap and rigorously derive upper bounds on the expected error for sparse grid least squares regression. Furthermore, we will verify that our theoretical convergence results also match the observed rates in numerical experiments.

Acknowledgements The author was supported by the Sonderforschungsbereich 1060 *The Mathematics of Emergent Effects* funded by the Deutsche Forschungsgemeinschaft.

1 Introduction

One of the most common tasks in *Big Data* applications is *function regression*. Here, we aim to approximate a function $g : \Omega \to \mathbb{R}$ defined on an open domain $\Omega \subset \mathbb{R}^m$. However, we only have access to n (possibly noisy) evaluations $(\mathbf{t}_i, g(\mathbf{t}_i) + \varepsilon_i) \in \Omega \times \mathbb{R}, i = 1, \ldots, n$ of g. Note that this is a special instance of a much more general regression or even density estimation problem, see e.g. [15].

Although many successful regression algorithms such as generalized clustering methods, radial basis function neural networks or support vector machines have been proposed over the last decades, see e.g. [1, 14, 20], one of the main problems in Big Data applications, namely the vast number n of data points, still presents a severe limitation to these so-called *data-centered* algorithms. This phenomenon usually

Bastian Bohn
Institute for Numerical Simulation, University of Bonn, Wegelerstr. 6, 53115 Bonn, e-mail: bohn@ins.uni-bonn.de

prevents the user from applying the above mentioned methods straightforwardly because of their superlinear runtime dependence on n, i.e. the number of computational steps grows much faster than n, e.g. $\mathscr{O}(n^3)$ for applying direct solvers to the regression problem. In order to cope with this problem, several enhancements to these algorithms, such as chunking or sparse greedy matrix approximation, have been introduced, see [20]. Furthermore, since many data-centered methods are based on *kernel representations*, we need to have access to a closed form of an appropriate kernel function. However, in many cases only infinite series expansion kernels are provided and an evaluation is not straightforward, see [12, 13].

To circumvent these issues and obtain an algorithm which naturally employs linear runtime complexity with respect to n, grid based discretizations have been proposed. Here, sparse grids are particularly well-suited since they allow to efficiently treat also higher-dimensional domains, i.e. $m > 3$, which is not possible with full tensor-product grids due to the *curse of dimensionality*. This means that - for a full grid space - the number of grid points N_k scales like $\mathscr{O}\left(2^{km}\right)$, where k denotes the grid level. In the sparse grid case, however, the scaling of N_k is only $\mathscr{O}\left(2^k k^{m-1}\right)$. Many variants of sparse grid regression algorithms can be found in e.g. [3, 5, 9, 10, 19].

Even though sparse grid regression algorithms have proven to be a good choice for many practical Big Data problems, there has not yet been a thorough theoretical justification for their good performance, i.e. the overall error convergence behavior and suitable couplings between N_k and n have yet to be determined. In this paper, we aim to close this gap for the case of (unregularized) least-squares function regression. Here, the corresponding problem is to determine

$$\arg\min_{h\in V_k} \frac{1}{n}\sum_{i=1}^{n}(h(\mathbf{t}_i)-g(\mathbf{t}_i)-\varepsilon_i)^2,$$

where V_k is the sparse grid space of level k, i.e. we search for the function $h \in V_k$ which minimizes the average squared distance between point evaluations of h and the unknown function g in the input data points $\mathbf{t}_i, i = 1,\ldots,n$. The evaluation in $\mathbf{t}_i$ is perturbed by some additive noise term ε_i. For this setting, we will derive the optimal coupling between N_k and n and present the corresponding error convergence rate. As we will see, the rate is governed mainly by the best approximation error in the sparse grid space and a sample-dependent term in which the noise variance σ will play an important role. To obtain our results, we will enhance the analysis of [7] on least-squares regression with orthonormal basis sets, which has been applied to derive convergence properties for global polynomial spaces in [6, 17, 18], to arbitrary basis sets and apply it to our sparse grid basis functions. While the choice of the particular basis is arbitrary in the orthonormal case, the quotient of the frame constants enters our estimates for non-orthonormal bases. Therefore, we use the sparse grid prewavelets since they form an L_2 Riesz frame and reveal essentially the same properties in our estimates as an orthonormal basis does in [7]. Furthermore, the prewavelets lead to sparsely populated system matrices for least-squares regression because of their compact support. Thus, our basis choice leads to a fast least-squares algorithm with quasi-optimal convergence rate in the piecewise linear case.

The remainder of this paper is structured as follows: In section 2 we recapitulate the least squares regression problem and introduce the necessary notation. Then, we briefly present our sparse grid spaces and the according basis functions in section 3. Our main results on the coupling and the convergence rate can be found in section 4. Subsequently, we provide numerical experiments to underscore our theoretical results in section 5. Finally, we conclude in section 6 with a short summary and an outlook on possible future research directions.

2 Least-squares regression

We now define the necessary ingredients to state and analyze the least squares function regression problem. To this end, let ρ be a probability measure on the Lebesgue σ-algebra of $\Omega \subset \mathbb{R}^m$ and let $g : \Omega \to \mathbb{R}$ be a point-evaluable, bounded function, i.e. there exists an $r > 0$ such that $\|g\|_{L_\infty,\rho(\Omega)} \leq r$. We define a real-valued random variable $\varepsilon = \varepsilon(\mathbf{t})$, which models the noise and fulfills

$$\mathbb{E}\left[\varepsilon \mid \mathbf{t}\right] = 0 \text{ for all } \mathbf{t} \in \Omega \quad \text{and} \quad \sigma^2 := \sup_{\mathbf{t}\in\Omega} \mathbb{E}\left[\varepsilon^2 \mid \mathbf{t}\right] < \infty. \tag{1}$$

Our n input data points for the least-squares regression are then given by

$$\mathscr{X}_n := (\mathbf{t}_i, g(\mathbf{t}_i) + \varepsilon_i)_{i=1}^n \subset \Omega \times \mathbb{R},$$

where the $\mathbf{t}_i$ are drawn i.i.d. according to ρ and the $\varepsilon_i = \varepsilon(\mathbf{t}_i)$ are instances of the random variable ε. Finally, we denote our scale of finite-dimensional search spaces, i.e. the spaces in which the solution to the regression problem will lie, by $V_k \subset L_{2,\rho}(\Omega)$ for a scale parameter $k \in \mathbb{N}$, which will be the level of our grid spaces later on. In the following we will write $N_k := \dim(V_k)$ to denote the dimension of the search space of level k. Then, as already mentioned in the introduction, we can write the least-squares regression problem as

$$\text{Determine } f_{\mathscr{X}_n,V_k} := \arg\min_{h\in V_k} \frac{1}{n} \sum_{i=1}^n \left(h(\mathbf{t}_i) - g(\mathbf{t}_i) - \varepsilon_i\right)^2. \tag{2}$$

Note that a regularized version of this problem, where a penalty term is added to the above formulation, is also often considered. However, in this paper we solely focus on the unregularized case (2) and give sufficient conditions such that this problem is stably solvable also without a penalty term.

To solve (2), let $v_1, \ldots, v_{N_k}$ be an arbitrary basis of V_k. Then it is straightforward to show that the coefficients $\alpha := \left(\alpha_1, \ldots, \alpha_{N_k}\right)^T$ of $f_{\mathscr{X}_n,V_k} = \sum_{i=1}^{N_k} \alpha_i v_i$ can be computed by solving the linear system

$$nBB^T\alpha = B\mathbf{x}, \tag{3}$$

where $B \in \mathbb{R}^{N_k \times n}$ is given by $B_{ij} := \frac{1}{n} v_i(\mathbf{t}_j)$ and $\mathbf{x} := (g(\mathbf{t}_1) + \varepsilon_1, \ldots, g(\mathbf{t}_n) + \varepsilon_n)^T$. For a more detailed discussion on this system, we refer to [2, 9].

3 Full grids and sparse grids

In order to solve (3) on a full grid space, i.e. $V_k = \mathscr{V}_k^{\text{full}}$ of level k, or a sparse grid space, i.e. $V_k = \mathscr{V}_k^{\text{sparse}}$ of level k, we have to define appropriate basis functions $v_1, \dots, v_{N_k}$. To this end, we consider the so-called piecewise linear prewavelet basis, see also [11], since it forms a Riesz frame, which will be of major importance for the analysis in the subsequent section. The prewavelets are based on linear combinations of the hat functions

$$\phi_{l,i}(t) := \phi(2^l \cdot t - i)|_{[0,1]} \quad \text{with} \quad \phi(t) := \begin{cases} 1 - |t| & \text{if } t \in [-1,1], \\ 0 & \text{else.} \end{cases} \tag{4}$$

The univariate prewavelet basis functions $\gamma_{l,i} : [0,1] \to \mathbb{R}$ are then defined by

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

for $l \leq 1$ and by

$$\gamma_{l,i} := 2^{\frac{l}{2}} \cdot \left(\frac{1}{10}\phi_{l,i-2} - \frac{6}{10}\phi_{l,i-1} + \phi_{l,i} - \frac{6}{10}\phi_{l,i+1} + \frac{1}{10}\phi_{l,i+2} \right)$$

for $l \geq 2$ and $i \in I_l \setminus \{1, 2^l - 1\}$ with $I_l := \left\{ i \in \mathbb{N} \mid 1 \leq i \leq 2^l - 1,\ i \text{ odd} \right\}$. For the boundary cases $i \in \{1, 2^l - 1\}$, we have

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

The m-variate prewavelet functions are defined by a simple product approach

$$\gamma_{\mathbf{l},\mathbf{i}}(\mathbf{t}) := \prod_{j=1}^{m} \gamma_{l_j,i_j}(t_j), \tag{5}$$

where $\mathbf{l} = (l_1, \dots, l_m)$ denotes the multivariate level index and $\mathbf{i} = (i_1, \dots, i_m)$ denotes the multivariate position index. The graph of two exemplary univariate and two exemplary bivariate prewavelet basis functions can be found in figure 1. In the multivariate case, the appropriate index sets are given by

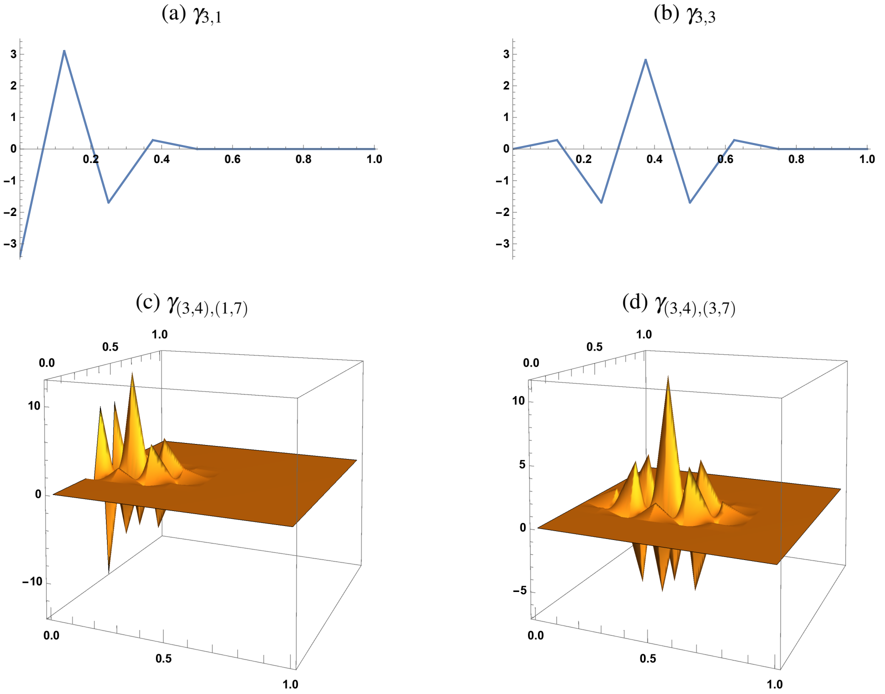
$$\mathbf{I}_{\mathbf{l}} := \left\{ \mathbf{i} \in \mathbb{N}^m \,\middle|\, \begin{matrix} 0 \leq i_j \leq 1, & \text{if } l_j = 0, \\ 1 \leq i_j \leq 2^{l_j} - 1,\ i_j \text{ odd} & \text{if } l_j > 0 \end{matrix} \ \text{for all } 1 \leq j \leq m \right\},$$

which lead to the hierarchical increment spaces

$$W_{\mathbf{l}} := \operatorname{span}\left\{ \gamma_{\mathbf{l},\mathbf{i}} \mid \mathbf{i} \in \mathbf{I}_{\mathbf{l}} \right\}.$$

Now, the full grid space of level $k > 0$ is defined by

Fig. 1: Piecewise linear prewavelet examples.



(a) $\gamma_{3,1}$

(b) $\gamma_{3,3}$

(c) $\gamma_{(3,4),(1,7)}$

(d) $\gamma_{(3,4),(3,7)}$

$$\mathcal{V}_k^{\text{full}} := \bigoplus_{\substack{\mathbf{l}\in\mathbb{N}^m \\ |\mathbf{l}|_{\ell_\infty}\le k}} W_{\mathbf{l}},$$

whereas the sparse grid space of level $k > 0$ is given by

$$\mathcal{V}_k^{\text{sparse}} := \bigoplus_{\substack{\mathbf{l}\in\mathbb{N}^m \\ \zeta_m(\mathbf{l})\le k}} W_{\mathbf{l}}$$

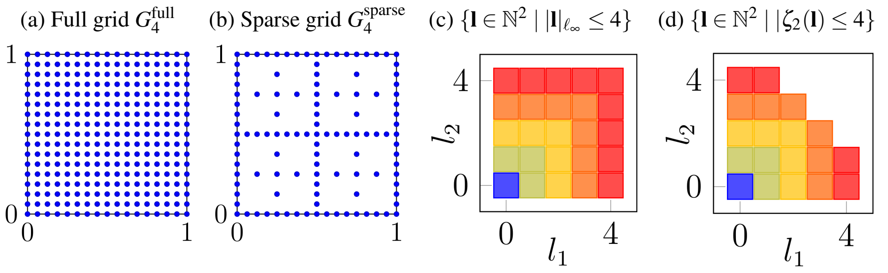
with $\zeta_m(\mathbf{0}) := 0$ and

$$\zeta_m(\mathbf{l}) := |\mathbf{l}|_{\ell_1} - m + \left|\{j \mid l_j = 0\}\right| + 1$$

for a non-zero $\mathbf{l} \in \mathbb{N}^m$. The specific choice of ζ_m guarantees that the highest resolution of a subgrid on the boundary is the same as the highest resolution of a subgrid in the interior of $[0,1]^m$. The corresponding grids G_k^{full} and G_k^{sparse}, i.e. the centers of the support of the involved prewavelet basis functions, can be found in figure 2.

As we mentioned above, full grids suffer from the curse of dimensionality, i.e. the degrees of freedom grow like

Fig. 2: Two-dimensional full grid and sparse grid and their corresponding index sets.


(a) Full grid G_4^{full}
(b) Sparse grid G_4^{sparse}
(c) $\{\mathbf{l} \in \mathbb{N}^2 \mid |\mathbf{l}|_{\ell_\infty} \leq 4\}$
(d) $\{\mathbf{l} \in \mathbb{N}^2 \mid |\zeta_2(\mathbf{l}) \leq 4\}$
0
1
0
1
0
1
0
1
4
0
l_2
0
4
l_1
4
0
l_2
0
4
l_1

$$\dim\left(\mathscr{V}_k^{\text{full}}\right) = (2^k+1)^m = \mathscr{O}\left(2^{km}\right),$$

which depends exponentially on the dimension m of the domain. For sparse grids, it can easily be obtained that

$$\dim\left(\mathscr{V}_k^{\text{sparse}}\right) = \mathscr{O}\left(2^k k^{m-1}\right)$$

see e.g. [4] for grids in the interior of the domain and [8] for grids which are also allowed to live on the boundary. As we see, the curse of dimensionality only appears with respect to the level k instead of 2^k. Therefore, sparse grids can be used also for $m > 3$.

4 Error analysis

After introducing the least-squares problem and our grid discretization in the previous sections, we can now present our main theorems on the stability and the error decay of a sparse grid regression algorithm. Our results are built on theorems 1 and 3 of [7] and can be seen as an extension thereof since only orthonormal bases are treated there, whereas our result also holds for arbitrary non-orthonormal bases.

4.1 Well-posedness and error decay

In the following, we denote the maximum and minimum eigenvalues of a symmetric matrix X by $\lambda_{\max}(X)$ and $\lambda_{\min}(X)$. We start with a Matrix Chernoff bound, which is proven in section 5 of [21].

Theorem 1 (Chernoff inequality for random matrices). *Let $D \in \mathbb{N}$ and $\delta \in [0,1)$ be arbitrary and let $X_1,\ldots,X_n \in \mathbb{R}^{D\times D}$ be independent, symmetric and positive* *semidefinite matrices with random entries. Let $R > 0$ be such that $\lambda_{\max}(X_i) \leq R$ holds for all $i = 1, \ldots, n$. Then, it holds*

$$\mathbb{P}\left[\lambda_{\min}\left(\sum_{i=1}^{n} X_i\right) \leq (1-\delta)c_{\min}\right] \leq D\left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\frac{c_{\min}}{R}}$$

and

$$\mathbb{P}\left[\lambda_{\max}\left(\sum_{i=1}^{n} X_i\right) \geq (1+\delta)c_{\max}\right] \leq D\left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\frac{c_{\max}}{R}}$$

with $c_{\min} := \lambda_{\min}\left(\mathbb{E}\left[\sum_{i=1}^{n} X_i\right]\right)$ and $c_{\max} := \lambda_{\max}\left(\mathbb{E}\left[\sum_{i=1}^{n} X_i\right]\right)$.

For a basis $v_1, \ldots, v_{N_k}$ of V_k, we introduce the quantity

$$S(v_1, \ldots, v_{N_k}) := \sup_{\mathbf{t}\in\Omega} \sum_{i=1}^{N_k} |v_i(\mathbf{t})|^2, \tag{6}$$

which will play a pivotal role throughout the rest of this paper. Note that this quantity is named $K(N_k)$ in [7] since it is independent of the basis choice there as the authors only deal with orthonormal bases. However, in our more general case, the quantity $S(v_1, \ldots, v_{N_k})$ is highly dependent on the concrete choice of the basis of V_k.

In the following, we denote the mass matrix on level k by $M = M(v_1, \ldots, v_{N_k}) \in \mathbb{R}^{N_k \times N_k}$, i.e. $M_{ij} = \langle v_i, v_j \rangle_{L_{2,\rho}(\Omega)}$. With the help of theorem 1, we are able to prove the following stability result, which is an extension of theorem 1 of [7].

Theorem 2 (Well-posedness). *Let $n \geq N_k$, $c = \left|\log\left(\frac{e^{0.5}}{(1.5)^{1.5}}\right)\right| \approx 0.1082$ and let*

$$S(v_1, \ldots, v_{N_k}) \leq c \cdot \frac{\lambda_{\min}(M)}{1+\theta} \cdot \frac{n}{\log(n)} \tag{7}$$

for a $\theta > 0$. Then, the solution $f_{\mathscr{X}_n,V_k} = \sum_{j=1}^{N_k} \alpha_j v_j$ of (3) exists, is unique and fulfills

$$\|f_{\mathscr{X}_n,V_k}\|_{L_{2,\rho}(\Omega)} \leq \sqrt{6} \cdot \frac{\lambda_{\max}(M)}{\lambda_{\min}(M)} \cdot \frac{1}{\sqrt{n}} \|\mathbf{x}\|_{\ell_2}$$

with probability at least $1 - 2n^{-\theta}$, where $\mathbf{x} := (g(\mathbf{t}_1) + \varepsilon_1, \ldots, g(\mathbf{t}_n) + \varepsilon_n)^T$.

Proof. The proof follows the lines of [7] with the necessary generalizations for arbitrary basis functions. Let $X \in \mathbb{R}^{N_k \times N_k}$ be the random, positive semi-definite matrix with entries $X_{ij} := \frac{1}{n} v_i(\mathbf{t}) \cdot v_j(\mathbf{t})$, where $\mathbf{t}$ is drawn according to ρ and let $X_1, \ldots, X_n$ be n realizations of X with $\mathbf{t} = \mathbf{t}_1, \ldots, \mathbf{t}_n$ from the samples $\mathscr{X}_n$. Then, $nBB^T = \sum_{i=1}^{n} X_i$ and $M = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]$.

Note that $\lambda_{\max}(X) \leq \frac{1}{n} S(v_1, \ldots, v_{N_k})$ almost surely since $X = nAA^T$ with $A = \frac{1}{n}\left(v_1(\mathbf{t}), \ldots, v_{N_k}(\mathbf{t})\right)^T$ and we have

$$\begin{aligned}\lambda_{\max}(X) &= n\lambda_{\max}(AA^T) = n\cdot\max_{|y|=1}\|Ay\|_{\ell_2}^2 = n\cdot\left(\frac{1}{n^2}\sum_{i=1}^{N_k}|\nu_i(\mathbf{t})\cdot 1|^2\right)\\ &\le \frac{1}{n}S(\nu_1,\dots,\nu_{N_k}).\end{aligned}$$

Therefore, we can apply theorem 1 with $D=N_k$, $R=\frac{1}{n}S(\nu_1,\dots,\nu_{N_k})$ and $\delta=\frac{1}{2}$ to obtain

$$\begin{aligned}P &:= \mathbb{P}\left[\lambda_{\min}(nBB^T)\le\frac{\lambda_{\min}(M)}{2}\quad\text{or}\quad\lambda_{\max}(nBB^T)\ge\frac{3\lambda_{\max}(M)}{2}\right]\\ &\le N_k\left(\frac{e^{-0.5}}{0.5^{0.5}}\right)^{\frac{n\lambda_{\min}(M)}{S(\nu_1,\dots,\nu_{N_k})}} + N_k\left(\frac{e^{0.5}}{1.5^{1.5}}\right)^{\frac{n\lambda_{\max}(M)}{S(\nu_1,\dots,\nu_{N_k})}} \le 2N_k\left(\frac{e^{0.5}}{1.5^{1.5}}\right)^{\frac{n\lambda_{\min}(M)}{S(\nu_1,\dots,\nu_{N_k})}},\end{aligned}$$

where the last inequality follows from $\lambda_{\min}(M)\le\lambda_{\max}(M)$ and $0<\frac{e^{-0.5}}{0.5^{0.5}}<\frac{e^{0.5}}{1.5^{1.5}}<1$. Using (7) and the definition of c, we obtain

$$P \le 2N_k e^{-\frac{cn\lambda_{\min}(M)}{S(\nu_1,\dots,\nu_{N_k})}} \le 2N_k\cdot n^{-(1+\theta)} \le 2n^{-\theta}$$

since we assumed $N_k\le n$. Therefore, (3) is uniquely solvable with probability at least $1-2n^{-\theta}$. Noting that $\|B\|_{\mathrm{Lin}(\mathbb{R}^n,\mathbb{R}^{N_k})}^2 = \frac{1}{n}\|nBB^T\|_{\mathrm{Lin}(\mathbb{R}^{N_k},\mathbb{R}^{N_k})} = \frac{1}{n}\lambda_{\max}(nBB^T)$ holds for the operator norm of the linear operator B and writing the L_2 norm with the help of the mass matrix, we finally get

$$\begin{aligned}\|f_{\mathscr{X}_n,V_k}\|_{L_{2,\rho}(\Omega)}^2 &= \alpha^T M\alpha \overset{(3)}{=} \mathbf{x}^T B^T(nBB^T)^{-1}M(nBB^T)^{-1}B\mathbf{x}\\ &\le \|\mathbf{x}\|_{\ell_2}^2\|B\|_{\mathrm{Lin}(\mathbb{R}^n,\mathbb{R}^{N_k})}^2\lambda_{\max}((nBB^T)^{-1})^2\lambda_{\max}(M)\\ &= \frac{1}{n}\|\mathbf{x}\|_{\ell_2}^2\lambda_{\max}(nBB^T)\frac{1}{\lambda_{\min}(nBB^T)^2}\lambda_{\max}(M)\\ &\le \frac{1}{n}\|\mathbf{x}\|_{\ell_2}^2\frac{3\lambda_{\max}(M)}{2}\frac{4}{\lambda_{\min}(M)^2}\lambda_{\max}(M) = 6\frac{\lambda_{\max}(M)^2}{\lambda_{\min}(M)^2}\cdot\frac{1}{n}\|\mathbf{x}\|_{\ell_2}^2\end{aligned}$$

with probability at least $1-2n^{-\theta}$, which proves our assertion. □

Theorem 2 tells us that the regression problem with basis $\nu_1,\dots,\nu_{N_k}$ is stably solvable for all $k\in\mathbb{N}$ with high probability if the number of samples n is large enough such that $n\ge N_k$ and (7) are fulfilled and if the fraction $\frac{\lambda_{\max}(M)}{\lambda_{\min}(M)}$, i.e. the condition number of the mass matrix, does not grow too fast with $k\to\infty$. Note that it is also possible to prove a more general version of this theorem if a (Tikhonov) regularization term is added, see [2].

Recall the L_∞ bound r on the function g from which the data $\mathscr{X}_n$ is sampled. For our error bound, we need to define the truncation operator $\tau_r : L_{\infty,\rho}(\Omega)\to L_{\infty,\rho}(\Omega)$ by $\tau_r(f)(\cdot) := P_r(f(\cdot))$, where the convex projection $P_r:\mathbb{R}\to\mathbb{R}$ is defined by

$$P_r(x) = \begin{cases} x & \text{if } |x| \leq r, \\ \frac{x}{|x|} \cdot r & \text{else.} \end{cases}$$

Note that τ_r is a non-expansive operator with respect to the $L_{2,\rho}(\Omega)$ norm, i.e. $\|\tau_r(f_1) - \tau_r(f_2)\|_{L_{2,\rho}(\Omega)} \leq \|f_1 - f_2\|_{L_{2,\rho}(\Omega)}$ for all $f_1, f_2 \in L_{\infty,\rho}(\Omega)$. Now, we can provide a theorem on the expected error behavior.

Theorem 3 (Expected regression error). *Let $n \geq N_k$ and let $f_{\mathscr{Z}_n,V_k}$ be the solution to* (3) *- or $f_{\mathscr{Z}_n,V_k} = 0$ if no unique solution to* (3) *exists. Let, furthermore, $S(\nu_1,\ldots,\nu_{N_k})$ and n fulfill* (7) *for a fixed $\theta > 0$ and for all $k \in \mathbb{N}$. Then,*

$$\mathbb{E}\left[\|\tau_r\left(f_{\mathscr{Z}_n,V_k}\right) - g\|^2_{L_{2,\rho}(\Omega)}\right] \leq \left(1 + \frac{8c\lambda_{\max}(M)}{(1+\theta)\lambda_{\min}(M)\log(n)}\right) \inf_{f \in V_k} \|f - g\|^2_{L_{2,\rho}(\Omega)} + 8r^2 n^{-\theta} + 8\sigma^2 \left(\frac{\lambda_{\max}(M)}{\lambda_{\min}(M)}\right)^2 \cdot \frac{N_k}{n} \tag{8}$$

with c from (7)*. Here, the expectation is taken with respect to the product measure $\rho^n := \rho \times \ldots \times \rho$.*

Proof. Again, the proof generalizes the one in [7], where only orthonormal bases are considered. In the following we will just write L_p for $L_{p,\rho}(\Omega)$ with $p \in [1,\infty]$. Let $\Omega^n = \Omega \times \ldots \times \Omega$ and let

$$\Omega^n_+ := \left\{(\mathbf{t}_1,\ldots,\mathbf{t}_n) \in \Omega^n \mid \lambda_{\max}(nBB^T) \leq \frac{3\lambda_{\max}(M)}{2} \text{ and } \lambda_{\min}(nBB^T) \geq \frac{\lambda_{\min}(M)}{2}\right\}$$

and let $\Omega^n_- := \Omega^n \setminus \Omega^n_+$. We have already shown in the proof of theorem 2 that $\mathbb{P}(\Omega^n_-) \leq 2n^{-\theta}$ since (7) holds. Let us denote $E := \mathbb{E}\left[\|\tau_r\left(f_{\mathscr{Z}_n,V_k}\right) - g\|^2_{L_2}\right]$. Since $|\tau_r(f)(\mathbf{t}) - g(\mathbf{t})| \leq |\tau_r(f)(\mathbf{t})| + |g(\mathbf{t})| \leq 2r$ holds for all $f \in L_\infty$ and almost every $\mathbf{t} \in \Omega$, we obtain

$$\begin{aligned} E &= \int_{\Omega^n_+} \|\tau_r\left(f_{\mathscr{Z}_n,V_k}\right) - g\|^2_{L_2} \,\mathrm{d}\rho^n + \int_{\Omega^n_-} \|\tau_r\left(f_{\mathscr{Z}_n,V_k}\right) - g\|^2_{L_2} \,\mathrm{d}\rho^n \\ &\leq \int_{\Omega^n_+} \|\tau_r\left(f_{\mathscr{Z}_n,V_k}\right) - g\|^2_{L_2} \,\mathrm{d}\rho^n + \int_{\Omega^n_-} 4r^2 \,\mathrm{d}\rho^n \\ &\leq \int_{\Omega^n_+} \|\tau_r\left(f_{\mathscr{Z}_n,V_k}\right) - g\|^2_{L_2} \,\mathrm{d}\rho^n + 8r^2 n^{-\theta} \\ &\leq \int_{\Omega^n_+} \|f_{\mathscr{Z}_n,V_k} - g\|^2_{L_2} \,\mathrm{d}\rho^n + 8r^2 n^{-\theta}, \end{aligned} \tag{9}$$

where the last inequality holds since τ_r is non-expansive and $g = \tau_r(g)$ holds almost everywhere.

Next, we define the projection $P^n_{V_k}$ onto V_k by

$$P^n_{V_k}(f) := \operatorname*{arg\,min}_{h \in V_k} \frac{1}{n} \sum_{i=1}^{n} \left(h(\mathbf{t}_i) - f(\mathbf{t}_i)\right)^2,$$

which is well-defined for point-evaluable functions f on Ω_+^n since the coefficients of $P_{V_k}^n(f)$ are given by (3) if we substitute the vector $\mathbf{x}$ by $(f(\mathbf{t}_1),\ldots,f(\mathbf{t}_n))^T$. Note that the coefficients of $f_{\mathscr{Z}_n,V_k}$ are given by $P_{V_k}^n(g+\varepsilon)$. Furthermore, we need the (standard) orthogonal L_2 projector P_{V_k} onto V_k. Obviously, it holds $P_{V_k}^n \circ P_{V_k} = P_{V_k}$. Therefore, we have

$$\begin{aligned}\|f_{\mathscr{Z}_n,V_k} - g\|_{L_2}^2 &= \|P_{V_k}^n(g+\varepsilon) - P_{V_k}^n \circ P_{V_k}(g) + P_{V_k}(g) - g\|_{L_2}^2 \\ &= \|P_{V_k}^n\left(g - P_{V_k}(g)\right) + P_{V_k}^n(\varepsilon)\|_{L_2}^2 + \|g - P_{V_k}(g)\|_{L_2}^2 \\ &\leq 2\|P_{V_k}^n\left(g - P_{V_k}(g)\right)\|_{L_2}^2 + 2\|P_{V_k}^n(\varepsilon)\|_{L_2}^2 + \|g - P_{V_k}(g)\|_{L_2}^2 \qquad (10)\end{aligned}$$

since $Id - P_{V_k}$ is L_2-orthogonal on V_k. To bound (9) from above, we will now deal with each of the three summands in (10) separately.

First, note that $P_{V_k}^n\left(g - P_{V_k}(g)\right) = \sum_{i=1}^{N_k} \beta_i \nu_i$ with $\beta = (\beta_1,\ldots,\beta_{N_k})^T$ given by $\beta = \left(nBB^T\right)^{-1}\xi$ with $\xi = B\mathbf{a}$ and $a_j = g(\mathbf{t}_j) - P_{V_k}(g)(\mathbf{t}_j)$ for $j = 1,\ldots,n$. Thus, we have

$$\begin{aligned}\|P_{V_k}^n\left(g - P_{V_k}(g)\right)\|_{L_2}^2 &= \beta^T M \beta = \xi^T \left(nBB^T\right)^{-1} M \left(nBB^T\right)^{-1} \xi \\ &\leq \lambda_{\max}(M) \frac{1}{\lambda_{\min}\left(nBB^T\right)^2} \|\xi\|_{\ell_2}^2 \leq \frac{4\lambda_{\max}(M)}{\lambda_{\min}(M)^2} \|\xi\|_{\ell_2}^2 \qquad (11)\end{aligned}$$

on Ω_+^n, on which nBB^T is invertible. This yields

$$\int_{\Omega_+^n} 2\|P_{V_k}^n\left(g - P_{V_k}(g)\right)\|_{L_2}^2 \,\mathrm{d}\rho^n \leq \frac{8\lambda_{\max}(M)}{\lambda_{\min}(M)^2} \mathbb{E}\left[\|\xi\|_{\ell_2}^2\right]. \qquad (12)$$

With the independence of $\mathbf{t}_1,\ldots,\mathbf{t}_n$, we deduce

$$\begin{aligned}\mathbb{E}\left[\|\xi\|_{\ell_2}^2\right] &= \int_{\Omega^n} \sum_{j=1}^{N_k} \left(\frac{1}{n}\sum_{i=1}^{n} \nu_j(\mathbf{t}_i) \cdot (g - P_{V_k}(g))(\mathbf{t}_i)\right)^2 \mathrm{d}\rho^n(\mathbf{t}_1,\ldots,\mathbf{t}_n) \\ &= \frac{1}{n^2}\sum_{j=1}^{N_k} (n^2 - n) \left(\underbrace{\int_{\Omega} \nu_j(\mathbf{t}) \cdot (g - P_{V_k}(g))(\mathbf{t})\,\mathrm{d}\rho(\mathbf{t})}_{=0}\right)^2 \\ &+ \frac{1}{n^2}\sum_{j=1}^{N_k} n \int_{\Omega} \left(\nu_j(\mathbf{t}) \cdot (g - P_{V_k}(g))(\mathbf{t})\right)^2 \mathrm{d}\rho(\mathbf{t}) \\ &\overset{(6)}{\leq} \frac{1}{n} S(\nu_1,\ldots,\nu_{N_k}) \|g - P_{V_k}(g)\|_{L_2}^2 \overset{(7)}{\leq} \frac{c\lambda_{\min}(M)}{(1+\theta)\log(n)} \|g - P_{V_k}(g)\|_{L_2}^2.\end{aligned}$$

Applying this to (12), we finally obtain

$$\int_{\Omega_+^n} 2\|P_{V_k}^n\left(g - P_{V_k}(g)\right)\|_{L_2}^2 \,\mathrm{d}\rho^n \leq \frac{8c\lambda_{\max}(M)}{(1+\theta)\lambda_{\min}(M)\log(n)} \|g - P_{V_k}(g)\|_{L_2}^2. \qquad (13)$$

For the second summand of (10), we proceed similarly. Note that $\vartheta = \left(nBB^T\right)^{-1}\eta$ are the coefficients of $P^n_{V_k}(\varepsilon)$ with respect to $\nu_1,\ldots,\nu_{N_k}$. Here, $\eta = B\mathbf{b}$ with $b_i = \varepsilon(\mathbf{t}_i)$. Analogously to (11), we get

$$\|P^n_{V_k}(\varepsilon)\|^2_{L_2} \leq \frac{4\lambda_{\max}(M)}{\lambda_{\min}(M)^2}\|\eta\|^2_{\ell_2}$$

on Ω^n_+. Therefore, it remains to estimate

$$\int_{\Omega^n_+} 2\|P^n_{V_k}(\varepsilon)\|^2_{L_2}\,\mathrm{d}\rho^n \leq \frac{8\lambda_{\max}(M)}{\lambda_{\min}(M)^2}\mathbb{E}\left[\|\eta\|^2_{\ell_2}\right]. \tag{14}$$

Because of (1) we have $\mathbb{E}_\rho[\varepsilon\nu_j] = 0$ for all $j \in 1,\ldots,N_k$. Thus, we obtain

$$\begin{aligned}
\mathbb{E}_{\rho^n}\left[\|\eta\|^2_{\ell_2}\right] &= \int_{\Omega^n}\sum_{j=1}^{N_k}\left(\frac{1}{n}\sum_{i=1}^{n}\nu_j(\mathbf{t}_i)\cdot\varepsilon(\mathbf{t}_i)\right)^2\mathrm{d}\rho^n(\mathbf{t}_1,\ldots,\mathbf{t}_n)\\
&= \frac{1}{n^2}\sum_{j=1}^{N_k}(n^2-n)\left(\underbrace{\mathbb{E}_\rho\left[\varepsilon\nu_j\right]}_{=0}\right)^2 + \frac{1}{n^2}\sum_{j=1}^{N_k}n\mathbb{E}_\rho\left[\varepsilon^2\nu_j^2\right]\\
&= \frac{1}{n}\sum_{j=1}^{N_k}\int_\Omega \nu_j(\mathbf{t})^2\mathbb{E}_\rho[\varepsilon^2\mid\mathbf{t}]\mathrm{d}\rho(\mathbf{t}) \overset{(1)}{\leq} \frac{\sigma^2}{n}\sum_{j=1}^{N_k}\int_\Omega \nu_j(\mathbf{t})^2\mathrm{d}\rho(\mathbf{t})\\
&\leq \frac{\sigma^2}{n}\sum_{j=1}^{N_k}\lambda_{\max}(M) = \frac{N_k\sigma^2}{n}\lambda_{\max}(M).
\end{aligned}$$

Plugging this into (14), we get

$$\int_{\Omega^n_+} 2\|P^n_{V_k}(\varepsilon)\|^2_{L_2}\,\mathrm{d}\rho^n \leq \frac{8\sigma^2\lambda_{\max}(M)^2}{\lambda_{\min}(M)^2}\cdot\frac{N_k}{n}. \tag{15}$$

Since the third summand of (10) is independent of the samples, we have

$$\int_{\Omega^n_+}\|g-P_{V_k}(g)\|^2_{L_2} \leq \|g-P_{V_k}(g)\|^2_{L_2} = \inf_{f\in V_k}\|g-f\|^2_{L_2}.$$

Finally, we combine this estimate together with (13) and (15) into (9) and (10), which completes the proof. □

The first term of the expected rate from theorem 3 depends mainly on the best approximation error in V_k and the quotient $\frac{\lambda_{\max}(M)}{\lambda_{\min}(M)}$, which can be bounded from above independently from k for Riesz bases for example. The second summand scales like $n^{-\theta}$, which resembles the decay of the error with respect to the amount of data n in the noiseless case, i.e. when $\sigma^2 = 0$ and the third summand vanishes. In the noisy case, the third summand is also present and the best possible decay rate with respect to n scales like n^{-1}.

4.2 Application to sparse grids

In the following, we assume that the measure ρ is the m-dimensional Lebesgue measure on $\Omega = [0,1]^m$, i.e. the data $\mathbf{t}_i, i = 1, \ldots, n$ are distributed uniformly in Ω. We now apply theorems 2 and 3 to the regression problem on sparse grid spaces $V_k = \mathscr{V}_k^{\text{sparse}}$ and need to bound

$$S(v_1, \ldots, v_{N_k}) = \sup_{\mathbf{t} \in \Omega} \sum_{\zeta_m(\mathbf{l}) \leq k} \sum_{\mathbf{i} \in \mathbf{I_l}} \gamma_{\mathbf{l},\mathbf{i}}(\mathbf{t})^2.$$

from above. To this end, we provide the following lemma.

Lemma 1. *For each* $\mathbf{l} \in \mathbb{N}^m$*, it holds*

$$\max_{\mathbf{t} \in [0,1]^m} \sum_{\mathbf{i} \in \mathbf{I_l}} \gamma_{\mathbf{l},\mathbf{i}}(\mathbf{t})^2 \leq 2^{|\mathbf{l}|_{\ell_1}} \cdot 2^{|\{j \in \{1,\ldots,m\} | l_j = 0\}|} \cdot \left(\frac{36}{25}\right)^{|\{j \in \{1,\ldots,m\} | l_j > 0\}|}. \qquad (16)$$

Proof. We first consider the univariate case $m = 1$ and define $S_l(t) := \sum_{i \in I_l} \gamma_{l,i}(t)^2$. For $l = 0$, we obtain

$$S_0(t) = \gamma_{0,0}^2(t) + \gamma_{0,1}^2(t) = 1 + t^2 \leq 2$$

and for $l = 1$ we have

$$S_1(t) = \gamma_{1,1}^2(t) = (2\phi_{1,1}(t) - 1)^2 \leq 1$$

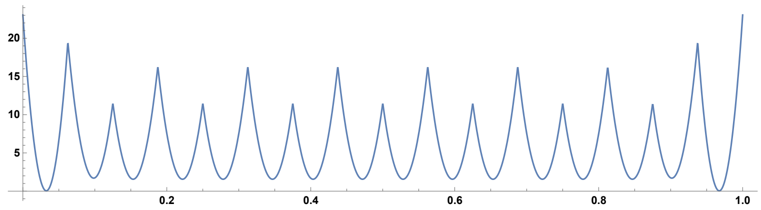
with $t \in [0,1]$. In the general case $l \geq 2$, S_l is a sum of the piecewise quadratic polynomials $\gamma_{l,i}^2(\cdot)$ with $i \in I_l$. Therefore, the quadratic term of the piecewise quadratic polynomial $S_l(\cdot)$ has a positive coefficient everywhere and the maximum of S_l over $[0,1]$ can only reside on one of the grid points $2^{-l}i$ with $i = 0, \ldots, 2^l$. This is also illustrated in figure 3, where S_4 is plotted exemplarily.

We now prove that the maximum of S_l is always attained at the boundary point $t = 1$. For $l = 0$ and $l = 1$, this is immediately clear. The (local) maxima of S_2 are denoted below in a mask-type notation which contains a prefactor 2^l and the nodal values at the grid points. The calculation

$$\begin{aligned} S_2(t) &= \gamma_{2,1}^2(t) + \gamma_{2,3}^2(t) \\ &= 4 \left[\ \tfrac{36}{25}\ \tfrac{121}{100}\ \tfrac{9}{25}\ \tfrac{1}{100}\ 0\ \right] \\ &+ 4 \left[\ 0\ \tfrac{1}{100}\ \tfrac{9}{25}\ \tfrac{121}{100}\ \tfrac{36}{25}\ \right] \\ &= 4 \left[\ \tfrac{36}{25}\ \tfrac{61}{50}\ \tfrac{18}{25}\ \tfrac{61}{50}\ \tfrac{36}{25}\ \right] \end{aligned}$$

shows that the largest value $4 \cdot \frac{36}{25}$ is attained at the boundary grid points. Analogously, we have

Fig. 3: The squared sum S_4 of the univariate prewavelet basis functions for $k = 4$.



$$\begin{aligned}
S_3(t) &= \gamma_{3,1}^2(t) + \gamma_{3,3}^2(t) + \gamma_{3,5}^2(t) + \gamma_{3,7}^2(t) \\
&= 8\,[\ \tfrac{36}{25}\ \tfrac{121}{100}\ \tfrac{9}{25}\ \tfrac{1}{100}\ 0\ 0\ 0\ 0\ 0\ ] \\
&+ 8\,[\ 0\ \tfrac{1}{100}\ \tfrac{9}{25}\ 1\ \tfrac{9}{25}\ \tfrac{1}{100}\ 0\ 0\ 0\ ] \\
&+ 8\,[\ 0\ 0\ 0\ \tfrac{1}{100}\ \tfrac{9}{25}\ 1\ \tfrac{9}{25}\ \tfrac{1}{100}\ 0\ ] \\
&+ 8\,[\ 0\ 0\ 0\ 0\ 0\ \tfrac{1}{100}\ \tfrac{9}{25}\ \tfrac{121}{100}\ \tfrac{36}{25}\ ] \\
&= 8\,[\ \tfrac{36}{25}\ \tfrac{61}{50}\ \tfrac{18}{25}\ \tfrac{51}{50}\ \tfrac{18}{25}\ \tfrac{51}{50}\ \tfrac{18}{25}\ \tfrac{61}{50}\ \tfrac{36}{25}\ ]
\end{aligned}$$

for $l = 3$. Due to the local support of the basis functions, analogous calculations show that the value of S_l never exceeds $2^l \cdot \frac{36}{25}$ also for higher levels l. Therefore, the maximum of S_l is always attained for $t = 1$. If $l = 0$, the maximum value is 2 and if $l \geq 2$, it is $2^l \cdot \frac{36}{25}$. For the special case $l = 1$, we use the crude estimate $S_1(1) = 1 < 2 \cdot \frac{36}{25}$. Therefore, the assertion (16) is proven for $m = 1$.

The case $m > 1$ follows directly from the tensor product construction of the basis. To see this, let $\mathbf{t} \in [0,1]^m$ and $\mathbf{l} \in \mathbb{N}^m$ be arbitrary. It holds

$$\sum_{\mathbf{i}\in\mathbf{I}_\mathbf{l}} \gamma_{\mathbf{l},\mathbf{i}}(\mathbf{t})^2 = \sum_{(i_1,\dots,i_m)\in\mathbf{I}_\mathbf{l}} \prod_{j=1}^m \gamma_{l_j,i_j}(t_j)^2 = \prod_{j=1}^m \sum_{i_j\in I_{l_j}} \gamma_{l_j,i_j}(t_j)^2$$

due to the structure of $\mathbf{I}_\mathbf{l}$. Therefore, the maximization of the term on the left can be split into the maximization of S_{l_j} for each direction $j \in \{1,\dots,m\}$. Since the maximum is bounded by 2 for directions j with $l_j = 0$ and by $2^{l_j} \cdot \frac{36}{25}$ for directions j with $l_j \geq 1$, the inequality (16) follows. □

We are now able to present an upper bound on $S(v_1,\dots,v_{N_k})$ for sparse grids.

Theorem 4. *For $V_k = \mathscr{V}_k^{sparse}$, $S(v_1,\dots,v_{N_k})$ can be bounded by*

$$S(v_1,\dots,v_{N_k}) \leq \left(\frac{72}{25}\right)^m (N_k + 1). \tag{17}$$

Proof. In the following, we write $Z(\mathbf{l}) := |\{j \in \{1,\dots,m\} \mid l_j = 0\}|$ for the number of zeros of a multiindex $\mathbf{l} \in \mathbb{N}^m$. Applying lemma 1, we obtain

$$S(v_1,\ldots,v_{N_k}) \leq \sum_{|\mathbf{l}|_{\ell_1}+Z(\mathbf{l})\leq k+m-1} 2^{|\mathbf{l}|_{\ell_1}+Z(\mathbf{l})} \cdot \left(\frac{36}{25}\right)^{m-Z(\mathbf{l})},$$

where we used $\zeta_m(\mathbf{l}) = |\mathbf{l}|_{\ell_1} - m + Z(\mathbf{l}) + 1$. Substituting $i = |\mathbf{l}|_{\ell_1} + Z(\mathbf{l})$, this becomes

$$S(v_1,\ldots,v_{N_k}) \leq \sum_{i=0}^{k+m-1} 2^i \cdot \sum_{l=0}^{m} |\{\mathbf{l}\in\mathbb{N}^m \mid |\mathbf{l}|_{\ell_1} = i-l \text{ and } Z(\mathbf{l}) = l\}| \cdot \left(\frac{36}{25}\right)^{m-l}.$$

Obviously, it holds $|\{\mathbf{l}\in\mathbb{N}^m \mid |\mathbf{l}|_{\ell_1} = i-l \text{ and } Z(\mathbf{l}) = l\}| = 0$ for all $l = 0,\ldots,m$ if $i < m$. Therefore, we can begin the summation over i from m. If $i \geq m$ holds, a simple combinatorial argument, see also [4], leads to

$$\begin{aligned}|\{\mathbf{l}\in\mathbb{N}^m \mid |\mathbf{l}|_{\ell_1} = i-l \text{ and } Z(\mathbf{l}) = l\}| &= |\{\mathbf{l}\in(\mathbb{N}\setminus\{0\})^{m-l} \mid |\mathbf{l}|_{\ell_1} = i-l\}| \cdot \binom{m}{l} \\ &= \binom{i-l-1}{m-l-1}\binom{m}{l}\end{aligned}$$

for arbitrary $l = 0,\ldots,m-1$ and furthermore

$$|\{\mathbf{l}\in\mathbb{N}^m \mid |\mathbf{l}|_{\ell_1} = i-m \text{ and } Z(\mathbf{l}) = m\}| = \begin{cases} 1 & \text{if } i = m \\ 0 & \text{else} \end{cases} = \delta_{im}.$$

Therefore, we have

$$\begin{aligned} S(v_1,\ldots,v_{N_k}) &\leq \sum_{i=m}^{k+m-1} 2^i \cdot \left(\delta_{im} + \sum_{l=0}^{m-1}\binom{i-l-1}{m-l-1}\binom{m}{l}\left(\frac{36}{25}\right)^{m-l}\right) \\ &= 2^m \cdot \sum_{i=0}^{k-1} 2^i \cdot \left(\delta_{i0} + \sum_{l=0}^{m-1}\binom{i+m-l-1}{m-l-1}\binom{m}{l}\left(\frac{36}{25}\right)^{m-l}\right) \\ &= 2^m \cdot \left(1 + \sum_{l=0}^{m-1}\left(\frac{36}{25}\right)^{m-l}\binom{m}{l}\left(\sum_{i=0}^{k-1} 2^i \binom{i+m-l-1}{m-l-1}\right)\right) \\ &= 2^m + 2^m \sum_{l=0}^{m-1}\left(\frac{36}{25}\right)^{m-l}\binom{m}{l}|G_k^{m-l}|, \end{aligned}$$

where $|G_k^{m-l}|$ denotes the size of an $m-l$-dimensional level-k sparse grid without boundary, see lemma 3.6 of [4] for a proof. To derive a bound with respect to the number of grid points N_k in a sparse grid with boundary points of level k in dimension m, we rewrite the above inequality by

$$\begin{aligned} S(v_1,\ldots,v_{N_k}) &\leq 2^m + \sum_{l=0}^{m-1}\left(2\cdot\frac{36}{25}\right)^{m-l} \cdot 2^l \binom{m}{l}|G_k^{m-l}| \\ &\leq 2^m + \left(\frac{72}{25}\right)^m \cdot \sum_{l=0}^{m-1} 2^l \binom{m}{l}|G_k^{m-l}| = 2^m + \left(\frac{72}{25}\right)^m N_k, \end{aligned}$$

where the last equality is proven in lemma 2.1.2 of [8]. Since $2 < \frac{72}{25} = 2.88$, this completes the proof. □

Combining the statements of theorem 2 and 4, we see that the sparse grid regression problem is well-posed with probability larger than $1 - 2n^{-\theta}$ if

$$\left(\frac{72}{25}\right)^m (N_k + 1) \leq c \frac{\lambda_{\min}(M)}{1+\theta} \cdot \frac{n}{\log(n)}. \tag{18}$$

Since the prewavelet basis of V_k is a Riesz frame with respect to the $L_{2,\rho}(\Omega)$ norm, the fraction $\frac{\lambda_{\max}(M)}{\lambda_{\min}(M)}$ is bounded from above independently of the level $k \in \mathbb{N}$. Therefore, the necessary scaling is essentially

$$N_k \simeq 2^k k^{m-1} \lesssim \frac{n}{\log(n)},$$

where the $\lesssim$ notation implies an m- and θ-dependent constant. The following corollary states our main result for sparse grids. There we deal with the (Bessel-potential) Sobolev spaces $H^s_{\rho,\mathrm{mix}}(\Omega)$ of dominating mixed smoothness with respect to the $L_{2,\rho}(\Omega)$ measure, see e.g. [2, 16].

Corollary 1 (Regression error for sparse grids). *Let $g \in H^s_{\rho,mix}(\Omega)$ for some $0 < s \leq 2$ and let $V_k = \mathcal{V}_k^{sparse}$. Let, furthermore,* (18) *hold for an arbitrary $\theta > 0$. Then, the regression problem is well-posed in the sense of theorem 2 with probability at least $1 - 2n^{-\theta}$ and the expected error fulfills*

$$\mathbb{E}\left[\|\tau_r\left(f_{\mathscr{X}_n,V_k}\right) - g\|^2_{L_{2,\rho}(\Omega)}\right] \leq C_{m,s,\theta,\sigma}\left(2^{-2sk}k^{m-1} + \frac{1}{n^\theta} + \frac{2^k k^{m-1}}{n}\right) \tag{19}$$

with a constant $C_{m,s,\theta,\sigma}$, which depends on m, s, θ, σ and $\|g\|_{H^s_{\rho,mix}(\Omega)}$.

Proof. To prove the expected error, we combine theorems 3 and 4 and use that the squared best approximation error behaves like

$$\inf_{f \in \mathcal{V}_k^{\text{sparse}}} \|f - g\|^2_{L_{2,\rho}(\Omega)} \leq C_{m,s} 2^{-2sk} k^{m-1} \|g\|^2_{H^s_{\rho,\mathrm{mix}}(\Omega)}$$

for $g \in H^s_{\rho,\mathrm{mix}}(\Omega)$ with an m- and s-dependent constant $C_{m,s}$, see e.g. theorem 3.25 of [2]. Furthermore, $\frac{\lambda_{\max}(M)}{\lambda_{\min}(M)}$ is bounded from above independently of k since the prewavelet basis is a Riesz frame with respect to the $L_{2,\rho}(\Omega)$ norm. Together with the fact that $N_k \leq C_m 2^k k^{m-1}$ holds for an m-dependent constant C_m, see e.g. [8], the statement of the corollary follows immediately. □

Finally, we can ask for the optimal coupling between the number of samples n and the number of sparse grid basis functions N_k, which achieves the best possible convergence rate in the sense that the terms in the error estimate (19) are (approximately) balanced. The resulting coupling is stated in the following corollary.

Corollary 2 (Optimal coupling and convergence rate for sparse grids). *Let $g \in H^s_{\rho,mix}(\Omega)$ for some $0 < s \leq 2$ and let $V_k = \mathscr{V}_k^{sparse}$. Then, the following holds:*

1. *Let $\sigma^2 > 0$ (noisy case) and let* (18) *hold for a $\theta \geq \frac{2s}{2s+1}$. Then, the asymptotically optimal coupling between n and N_k is*

$$N_k \sim n^{\frac{1}{2s+1}} \log(n)^{m-1} \tag{20}$$

and the resulting convergence rate for $n \to \infty$ is

$$\mathbb{E}\left[\|\tau_r\left(f_{\mathscr{Z}_n,V_k}\right) - g\|^2_{L_{2,\rho}(\Omega)}\right] = \mathscr{O}\left(n^{-\frac{2s}{2s+1}} \log(n)^{m-1}\right). \tag{21}$$

2. *Let $\sigma^2 = 0$ (noiseless case) and let* (18) *hold for a $\theta > 2s$. Then, the asymptotically optimal coupling between n and N_k is*

$$N_k \sim \frac{n}{\log(n)} \tag{22}$$

and the resulting convergence rate for $n \to \infty$ is

$$\mathbb{E}\left[\|\tau_r\left(f_{\mathscr{Z}_n,V_k}\right) - g\|^2_{L_{2,\rho}(\Omega)}\right] = \mathscr{O}\left(n^{-2s} \log(n)^{(2s+1)m-1}\right). \tag{23}$$

Proof. Let $E := \mathbb{E}\left[\|\tau_r\left(f_{\mathscr{Z}_n,V_k}\right) - g\|^2_{L_{2,\rho}(\Omega)}\right]$. Note that $N_k \sim 2^k k^{m-1}$ in the sense that there exist two constants $c_1, c_2 > 0$ such that $c_1 2^k k^{m-1} \leq N_k \leq c_2 2^k k^{m-1}$ holds independently of k. Note, furthermore, that there exists a constants $C_1, C_2 > 0$ such that $C_1 \log(n) \leq k \leq C_2 \log(n)$ for $n \geq 2$ for each of the scalings (20) and (22). This can easily be obtained by taking the logarithm on both sides of (20) and (22).

We begin with the proof for the noisy case $\sigma^2 > 0$ and insert the coupling (20) into the error formula (19). Since we will see that this balances the first and third summands there, the coupling is also optimal. Indeed, we have

$$\begin{aligned} E \quad &\lesssim \quad 2^{-2sk} k^{m-1} + \frac{1}{n^\theta} + \frac{2^k k^{m-1}}{n} \lesssim (N_k)^{-2s} k^{(m-1)(2s+1)} + n^{-\theta} + \frac{N_k}{n} \\ &\lesssim \quad \left(n^{\frac{1}{2s+1}} \log(n)^{m-1}\right)^{-2s} \log(n)^{(m-1)(2s+1)} + n^{-\theta} + \frac{n^{\frac{1}{2s+1}} \log(n)^{m-1}}{n} \\ &\overset{\theta \geq \frac{2s}{2s+1}}{\lesssim} \quad n^{-\frac{2s}{2s+1}} \left(\log(n)^{m-1} + 1 + \log(n)^{m-1}\right) = \mathscr{O}\left(n^{-\frac{2s}{2s+1}} \log(n)^{m-1}\right). \end{aligned} \tag{24}$$

As we see in (24), the first and third summand of the error estimate (19) are balanced for the coupling (20). Note that the coupling is valid in the sense that it (asymptotically) fulfills condition (18). This completes the proof for the noisy case.

In the noiseless case $\sigma^2 = 0$, the third summand in (19) vanishes, see also theorem 3. Therefore, for $\theta = 2s + \delta$ with some arbitrary $\delta > 0$, the number of basis functions N_k needs to be chosen as large as possible (with respect to n) to achieve the fastest possible convergence of the first summand of (19). This is achieved by choosing n as the smallest integer such that (18) is still fulfilled, i.e. the corresponding scaling is (22). Therefore, we obtain

$$\begin{aligned} E &\lesssim 2^{-2sk}k^{m-1} + \frac{1}{n^{\theta}} \lesssim (N_k)^{-2s}\, k^{(m-1)(2s+1)} + n^{-\theta} \\ &\lesssim \left(\frac{n}{\log(n)}\right)^{-2s} \log(n)^{(m-1)(2s+1)} + n^{-\theta} \overset{\theta=2s+\delta}{\lesssim} n^{-2s}\left(\log(n)^{(2s+1)m-1} + n^{-\delta}\right) \\ &= \mathscr{O}\left(n^{-2s}\log(n)^{(2s+1)m-1}\right), \end{aligned}$$

which concludes the proof. □

For all of our proven convergence results, we see that the curse of dimensionality appears only in terms which scale logarithmically in the number of samples n. This is the well-known sparse grid effect, which we are used to when considering the spaces $\mathscr{V}_k^{\text{sparse}}$ for interpolation or approximation for instance, see [4].

As we see from corollary 2, the optimal main rate that can be achieved in the noisy case is $n^{-\frac{2s}{2s+1}}$, which becomes $n^{-\frac{4}{5}}$ in the smoothest setting ($s = 2$) that the piecewise linear basis functions can exploit.[1] This comes at an expense of oversampling by $n \sim N_k^{2s+1}$ if we neglect the logarithm. In the noiseless case, however, the much better main rate n^{-2s} can be achieved and there is only a logarithmic oversampling, see (22). This oversampling has to be present to fulfill the necessary condition (18) anyway.

Finally, note that our stability and error analysis for sparse grids heavily relies on the fact that we are dealing with a Riesz basis. Nevertheless, if we choose a basis for which $\frac{\lambda_{\max}(M)}{\lambda_{\min}(M)}$ is unbounded, e.g. the hierarchical hat basis built from $\phi_{l,i}$, see (4), we can still obtain well-posedness of the regression problem if an appropriate regularization term is added to (2), see also [2]. However, then it is not directly clear how to derive a variant of theorem 3 for the regularized case.

5 Numerical experiments

In this section, we have a look at numerical experiments, which illustrate our theoretical results from the previous section. To this end, we choose $\Omega = [0,1]^2$, $V_k = \mathscr{V}_k^{\text{sparse}}$ and $\rho = \lambda_{[0,1]^2}$ as the two-dimensional Lebesgue measure. We use the example function $g : [0,1]^2 \to \mathbb{R}$ given by

$$g(t_1,t_2) = \exp(-t_1^2 - t_2^2) + t_1 t_2. \tag{25}$$

Since g is infinitely smooth, we have $g \in H^2_{\rho,\text{mix}}((0,1)^2)$ and we can expect our results from the previous section to hold with smoothness index $s = 2$. We now discern two

[1] For higher order spline bases, a larger choice of s can be exploited here. However, one needs to prove an analogous result to theorem 4 for the corresponding basis functions first.

cases: The noiseless case, in which our samples are given as $\mathscr{Z}_n = (\mathbf{t}_i, g(\mathbf{t}_i))_{i=1}^n$, and the noisy case, where we deal with $\mathscr{Z}_n = (\mathbf{t}_i, g(\mathbf{t}_i) + \varepsilon_i)_{i=1}^n$ and the ε_i are independent instances of a normally distributed random variable $\varepsilon \sim \mathscr{N}(0, 0.01)$.

Since $\|g\|_{L_\infty,\rho([0,1]^2)} < 2$ and $\mathbb{P}[|\varepsilon| > 1] < 10^{-2000}$, we can safely assume that $r = 3$ is large enough to assure that (with probability almost 1) $|g(\mathbf{t}_i) + \varepsilon_i| < r$ holds for each $i =, 1 \ldots, n$. Therefore, $\tau_r\left(f_{\mathscr{Z}_n,V_k}\right) = f_{\mathscr{Z}_n,V_k}$ since $f_{\mathscr{Z}_n,V_k}$ is the optimal piecewise linear regression function in V_k and, thus, cannot be larger than $\max_{i=1,\ldots,n} |g(\mathbf{t}_i) + \varepsilon_i|$ anywhere. Therefore, we can apply theorem 3 and corollaries 1 and 2 for $f_{\mathscr{Z}_n,V_k}$ instead of $\tau_r(f_{\mathscr{Z}_n,V_k})$ in our setting.

Since the prewavelet basis is a Riesz frame, we know that $\frac{\lambda_{\max}(M)}{\lambda_{\min}(M)}$ is bounded independently of k. To see that this quotient is not severely large, we exemplarily calculated it for $k = 1, \ldots, 8$ and observed that it does not exceed 5 in the two-dimensional case.

Error decay

First, we compute the error for different pairs of grid levels k and numbers of data points n. Since our result on the regression error in corollary 1 is only given in expectation, we compute the average AvErr of the error $\|f_{\mathscr{Z}_n,V_k} - g\|_{L_{2,\rho}(\Omega)}^2$ over 10 independent runs with different input data sets for each parameter pair (k, n). To compute the error values, we interpolated both $f_{\mathscr{Z}_n,V_k}$ and g on a full tensor-product grid of level 11, i.e. we interpolated in $\mathscr{V}_{11}^{\text{full}}$, and computed the norm of the difference there. The results can be found in figure 4.

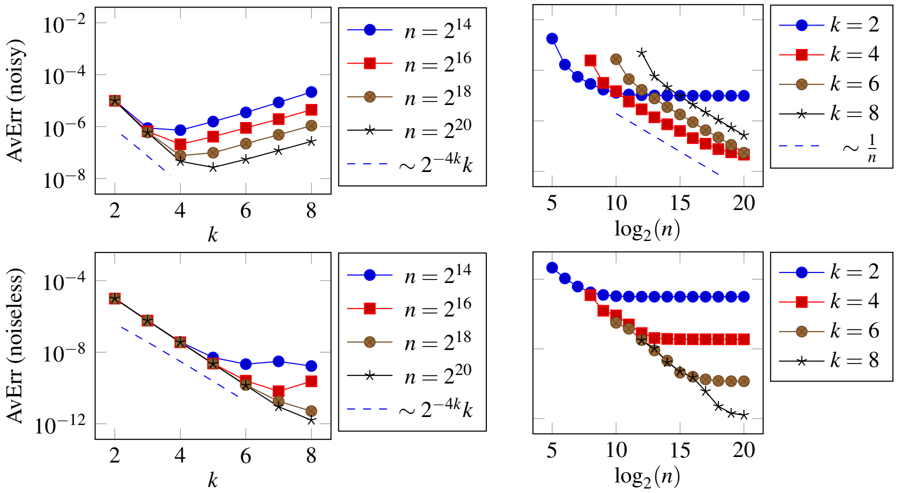
We directly observe the expected error decay rates, i.e. $2^{-4k} \cdot k$ for fixed n and n^{-1} for fixed k in the noisy setting (if we tacitly assume $\theta \geq 1$), see also corollary 1. For fixed k, we would expect the error to behave like $n^{-\theta}$ in the noiseless setting. However, since θ grows when the quotient $\frac{n}{k}$ grows, we cannot expect the error behavior to be of type n^{-p} for some p. For both, the noisy and the noiseless case, we observe that if the varying parameter (e.g. n) is too large, the error is saturated and the other parameter (e.g. k) has to be increased to guarantee a further error reduction. Note that the error for fixed n in the noisy regression setting even increases for large k. This is an overfitting effect, i.e. the basis size N_k is too large for the corresponding number of data n. Since there is no regularization in our approach, the error thus grows for large k and small n.

Balancing the error

In a next step, we balance the error terms according to corollary 2 and inspect the resulting convergence rates. For the noisy setting, we have for $\theta \geq \frac{4}{5}$ that the optimal coupling is given by

$$N_k \sim n^{\frac{1}{5}} \log(n).$$

Fig. 4: The average of $\|f_{\mathscr{Z}_n,V_k} - g\|^2_{L_{2,\rho}(\Omega)}$ over 10 runs for several parameter pairs (k,n) for which $N_k \leq n$ holds. Top: Noisy data, Bottom: Noiseless data. Left: Each line represents a fixed n, Right: Each line represents a fixed k.



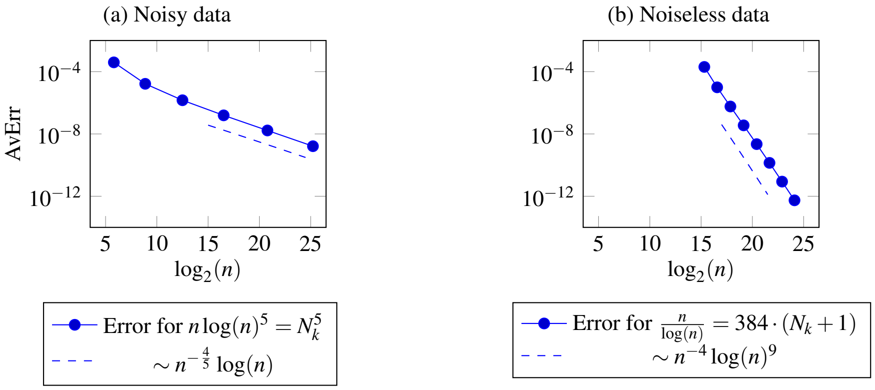
We, therefore, (approximately) solve $N_k^5 = n\log(n)^5$ for n and determine the optimal number of data points for $k = 1,\ldots,6$. For $k = 6$, the amount n of data points already exceeds 2^{25}. In the noiseless setting, the picture is quite different. Here, the optimal coupling is given by

$$N_k \sim \frac{n}{\log(n)}$$

if $\theta > 4$. More accurately, we look for the smallest n such that (18) is fulfilled with $\theta > 4$. Therefore, we equate both sides of (18) and (approximately) solve for n. Here, we set $\theta = 4$ and $\lambda_{\min}(M) = 1$ and obtain that sampling by $\frac{n}{\log(n)} = 384 \cdot (N_k + 1)$ suffices to fulfill (18). The average errors (over 10 runs) for the optimal coupling in the noisy and in the noiseless setting can be found in figure 5.

We directly see that the convergence rate in the experimental results asymptotically matches the proven rates from corollary 2, i.e. $n^{-\frac{2s}{2s+1}}\log(n)^{m-1} = n^{-\frac{4}{5}}\log(n)$ in the noisy case and $n^{-2s}\log(n)^{(2s+1)m-1} = n^{-4}\log(n)^9$ in the noiseless case. Furthermore, we observe that the initial error decay for noisy data is better than the convergence rate suggests. This is due to the fact that the noise effects the convergence behavior only if the overall error is already smaller than a certain (noise) level. Note also that the oversampling factor 384 is the reason why we already have more than 2^{15} data points for the smallest level $k = 1$ in the noiseless case. However, since our sampling resembles only a sufficient condition to ensure well-posedness of the regression problem with high probability, a much smaller oversampling constant might also do the job for practical applications.

Fig. 5: The average of $\|f_{\mathscr{X}_n,V_k} - g\|^2_{L_{2,\rho}(\Omega)}$ over 10 runs for the optimal coupling between k and n. Left: Noisy data with coupling $n\log(n)^5 = N_k^5$, Right: Noiseless data with coupling $\frac{n}{\log(n)} = 384\cdot(N_k+1)$, which resembles (18) for our example.



(a) Noisy data

(b) Noiseless data

6 Conclusion

In this article we presented error bounds, stability results and optimal parameter couplings for the least-squares regression problem and applied them to the sparse grid setting. To this end, we extended the results of [7] to arbitrary bases and provided an upper bound for the crucial quantity $S(v_1,\ldots,v_{N_k})$ from the stability and convergence estimates. Our results showed that the sparse grid prewavelet basis behaves (up to constants) like an orthonormal basis in the regression estimates because of its Riesz property. Therefore, it is a good choice for regression problems on sparse grid spaces since it employs both beneficial convergence behavior and small support of the corresponding basis functions, which is directly connected to the availability of cost-efficient linear equation system solvers, see e.g. [3, 5]. Finally, we presented a numerical example to illustrate that our results are not only of theoretical interest but resemble the true convergence behavior of actual sparse grid regression algorithms.

An interesting question which still has to be answered is if the general behavior of the growth of $S(v_1,\ldots,v_{N_k})$, see theorem 4, carries over also to higher-order spline bases on sparse grids. This is not directly clear from the proof techniques used in this paper as they rely on the piecewise linear structure of the regression function. Furthermore, it remains open how our results generalize to the regularized case, where a penalty term is added in the minimization problem. A first step into this direction regarding the stability estimate can be found in [2]. However, the rate of error decay and the optimal parameter coupling are still unknown in this case. Finally, a thorough comparison of our derived convergence rates for the sparse grid method with the error decay behavior of other regression algorithms such as support vector machines or multilayer neural networks still has to be done.

References

1. A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.
2. B. Bohn. *Error analysis of regularized and unregularized least-squares regression on discretized function spaces*. PhD thesis, Institute for Numerical Simulation, University of Bonn, 2017.
3. B. Bohn and M. Griebel. An adaptive sparse grid approach for time series predictions. In J. Garcke and M. Griebel, editors, *Sparse grids and applications*, volume 88 of *Lecture Notes in Computational Science and Engineering*, pages 1–30. Springer, 2012.
4. H.-J. Bungartz and M. Griebel. Sparse grids. *Acta Numerica*, 13:147–269, 2004.
5. H.-J. Bungartz, D. Pflüger, and S. Zimmer. Adaptive sparse grid techniques for data mining. In H. Bock, E. Kostina, X. Hoang, and R. Rannacher, editors, *Modelling, Simulation and Optimization of Complex Processes 2006, Proc. Int. Conf. HPSC, Hanoi, Vietnam*, pages 121–130. Springer, 2008.
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(3):815–837, 2015.
7. A. Cohen, M. Davenport, and D. Leviatan. On the stability and accuracy of least squares approximations. *Foundations of Computational Mathematics*, 13:819–834, 2013.
8. C. Feuersänger. *Sparse Grid Methods for Higher Dimensional Approximation*. PhD thesis, Institute for Numerical Simulation, University of Bonn, 2010.
9. J. Garcke. *Maschinelles Lernen durch Funktionsrekonstruktion mit verallgemeinerten dünnen Gittern*. PhD thesis, Institute for Numerical Simulation, University of Bonn, 2004.
10. J. Garcke, M. Griebel, and M. Thess. Data mining with sparse grids. *Computing*, 67(3):225–253, 2001.
11. M. Griebel and P. Oswald. Tensor product type subspace splitting and multilevel iterative methods for anisotropic problems. *Advances in Computational Mathematics*, 4:171–206, 1995.
12. M. Griebel, C. Rieger, and B. Zwicknagl. Multiscale approximation and reproducing kernel Hilbert space methods. *SIAM Journal on Numerical Analysis*, 53(2):852–873, 2015.
13. M. Griebel, C. Rieger, and B. Zwicknagl. Regularized kernel based reconstruction in generalized Besov spaces. *Accepted by Foundations of Computational Mathematics*, 2016.
14. T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.
15. M. Hegland. Data mining techniques. *Acta Numerica*, 10:313–355, 2001.
16. S. Knapek. *Approximation und Kompression mit Tensorprodukt-Multiskalenräumen*. PhD thesis, Institute for Numerical Simulation, University of Bonn, 2000.
17. G. Migliorati, F. Nobile, and R. Tempone. Convergence estimates in probability and in expectation for discrete least squares with noisy evaluations at random points. *Journal of Multivariate Analysis*, 142:167–182, 2015.
18. 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*, 14:419–456, 2014.
19. D. Pflüger, B. Peherstorfer, and H.-J. Bungartz. Spatially adaptive sparse grids for high-dimensional data-driven problems. *Journal of Complexity*, 26(5):508–522, 2010.
20. B. Schölkopf and A. Smola. *Learning with Kernels – Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press – Cambridge, Massachusetts, 2002.
21. J. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2011.