Weierstraß-Institut

für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Convergence bounds for empirical nonlinear least-squares

Martin Eigel[1], Philipp Trunschke[2], Reinhold Schneider[2]

submitted: April 14, 2020

[1] Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: martin.eigel@wias-berlin.de

[2] TU Berlin
Straße des 17. Juni 136
10623 Berlin
Germany
E-Mail: ptrunschke@mail.tu-berlin.de
schneidr@math.tu-berlin.de

No. 2714
Berlin 2020

2010 *Mathematics Subject Classification.* 41A10, 41A25, 41A65, 62E17, 93E24.

Key words and phrases. Multivariate approximation, Restricted isometry property, Weighted least squares, Tensor representation, Convergence rates, Error analysis, Nonlinear approximation, Conditional sampling.

M. Eigel acknowledges support by the DFG SPP 1886. R. Schneider was supported by the Einstein Foundation Berlin. P. Trunschke acknowledges support by the Berlin International Graduate School in Model and Simulation based Research (BIMoS). We would like to thank Henryk Gerlach for his help in generating 3D plots with `povray`. We also thank Leon Sallandt, Mathias Oster and Michael Götte for fruitful discussions.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Convergence bounds for empirical nonlinear least-squares

Martin Eigel, Philipp Trunschke, Reinhold Schneider

ABSTRACT. We consider best approximation problems in a nonlinear subset $\mathcal{M}$ of a Banach space of functions $(\mathcal{V}, \|\bullet\|)$. The norm is assumed to be a generalization of the L^2-norm for which only a weighted Monte Carlo estimate $\|\bullet\|_n$ can be computed. The objective is to obtain an approximation $v \in \mathcal{M}$ of an unknown function $u \in \mathcal{V}$ by minimizing the empirical norm $\|u - v\|_n$. In the case of linear subspaces $\mathcal{M}$ it is well-known that such least squares approximations can become inaccurate and unstable when the number of samples n is too close to the number of parameters $m = \dim(\mathcal{M})$. We review this statement for general nonlinear subsets and establish error bounds for the empirical best approximation error. Our results are based on a restricted isometry property (RIP) which holds in probability and we show that $n \gtrsim m$ is sufficient for the RIP to be satisfied with high probability. Several model classes are examined where analytical statements can be made about the RIP. Numerical experiments illustrate some of the obtained stability bounds.

1. INTRODUCTION, SCOPE, CONTRIBUTIONS

We consider the problem of estimating an unknown function u from noiseless observations. For this problem to be well-posed, some prior information about u has to be assumed, which often takes the form of regularity assumptions. To make this notion more precise, we assume that u is an element of some Banach space of functions $(\mathcal{V}, \|\bullet\|)$ that can be well approximated in a given (nonlinear) subset $\mathcal{M} \subseteq \mathcal{V}$. The approximation error is measured in the norm

$$\|v\| \coloneqq \left(\int_Y |v|_y^2 \,\mathrm{d}\rho(y)\right)^{1/2}$$

where Y is some Borel subset of $\mathbb{R}^d$, ρ is a probability measure on Y and $|\bullet|_y$ is a y-dependent seminorm for which the integral above is finite for all $v \in \mathcal{V}$. This norm is a generalization of the $L^2(Y,\rho)$- and $H^1_0(Y,\rho)$-norms which are induced by the seminorms $|v|_y^2 = |v(y)|^2$ and $|v|_y^2 = \|\nabla v(y)\|_2^2$, respectively.

We characterize the best approximation operator P implicitly by

$$Pu \in \operatorname*{arg\,min}_{v\in\mathcal{M}} \|u - v\|.$$

In general, this operator is not computable. We propose to approximate P by an estimator P_n that is based on the weighted least-squares method which replaces the norm $\|v\|$ by the empirical seminorm

$$\|v\|_n := \left(\frac{1}{n}\sum_{i=1}^n w(y_i)|v|_{y_i}^2\right)^{1/2}$$

for a given *weight function* w and a sample set $\{y_i\}_{i=1}^n \subseteq Y$ with $y_i \sim w^{-1}\rho$. The weight function is a non-negative function $w \geq 0$ such that $\int_Y w^{-1}\,\mathrm{d}\rho = 1$. The corresponding *empirical* best approximation operator P_n is characterized implicitly by

$$P_n u \in \operatorname*{arg\,min}_{v\in\mathcal{M}} \|u - v\|_n. \tag{1}$$

DOI 10.20347/WIAS.PREPRINT.2714 Berlin 2020

Given this definition, we can choose w such that an optimal rate of convergence $P_n \to P$ is achieved as n tends to ∞. Note that changing the sampling measure from ρ to $w^{-1}\rho$ is a common strategy to reduce the variance in Monte Carlo methods referred to as *importance sampling*.

Since $\|\bullet\|$ is not computable in general the best approximation error

$$\|(1-P)u\| = \min_{v\in\mathcal{M}} \|u-v\|$$

serves as a baseline for a numerical method founded on a finite set of samples. We prove in this paper that the empirical best approximation error $\|(1-P_n)u\|$ is equivalent to this error with high probability.

Main result. *Let $\mathcal{M} \subset \mathcal{V}$ be a model class and choose $\delta \in (0,1)$. There exist positive constants $K = K(u,\mathcal{M})$ and $\nu = \nu(u,\mathcal{M},\delta)$ such that*

$$\|(1-P)u\| \leq \|(1-P_n)u\| \leq \left(1 + 2\frac{\sqrt{1+\delta}}{\sqrt{1-\delta}}\right)\|(1-P)u\|$$

holds with probability $p \geq 1 - \nu\exp(-n\delta^2 K)$.

To put this in a broader perspective, we propose a convergence theory based on the restricted isometry property (RIP) as known from compressed sensing where only finite dimensional linear spaces are considered. The RIP can be conceived as a generalization of stability conditions such as the Ladyzhenskaya-Babuška-Brezzi (LBB) condition in finite element analysis. Our goal is an extension to nonlinear approximations and this paper provides first results towards a more general theory which can be used in particular in the context of low-rank tensor reconstruction as discussed in [1].

Our considerations are closely related to statistical learning as for instance presented in [2] and [3]. However, our samples can usually be actively generated and there is no intrinsic noise involved, although this can be introduced if required. This setting allows for results which are qualitatively much better than what can be achieved in learning theory.

Examples to which our theory applies are for instance finite dimensional vector spaces and sets of sparse vectors, or low-rank tensors. We later examine some model classes for which analytical statements regarding the RIP can be derived.

1.1. **Structure.** The remainder of the paper is organized as follows. In Section 1.2 we aim to provide a brief overview of previous work. Based on the notion of the RIP Section 2 develops the central results of this work. These are applied to some common model classes in Section 3. We begin by considering linear spaces in Section 3.1 and illustrate how the choice of the seminorm influences the convergence. Section 3.3 considers sets of sparse functions and Section 3.4 examines sets of low-rank functions. The connection to *empirical risk minimization (ERM)* as scrutinized in our earlier work [1] is discussed in Section 4. We conclude in Section 5 with a discussion of the derived results and an outlook on future work.

1.2. **Related work.** In statistics $P_n u$ is known as the nonlinear least squares estimator of u. The extensive use of machine learning in recent years has lead to the investigation of this estimator for special model classes like sparse vectors [4–6], low-rank tensors [1, 7, 8] and neural networks [9, 10]. However, to the knowledge of the authors no investigation for general model classes has been published so far.

The empirical approximation problem (1) was thoroughly examined in [11] for linear model spaces. There the model class $\mathcal{m}$ is assumed to be the m-dimensional subspace spanned by the *orthonormal* basis functions $\{\mathbf{B}_j\}_{j\in[m]}$ in $\mathcal{V} = L^2(Y,\rho)$. One of the key points in this paper is that for estimating the error of $\|(1-P_n)u\|$ it suffices to ensure that $\|\mathbf{G}-\mathbf{I}_m\|_2 \leq \delta < 1$ where $\mathbf{G} := \frac{1}{n}\sum_{i=1}^n w(y_i)\mathbf{B}(y_i)\mathbf{B}(y_i)^\intercal$ is the Monte Carlo estimate of the Gram matrix $\mathbf{I}_m$. This condition is in fact equivalent to the restricted isometry property (RIP)

$$(1-\delta)\|u\|^2 \leq \|u\|_n^2 \leq (1+\delta)\|u\|^2 \qquad \text{for } u \in \operatorname{span}(\boldsymbol{B}). \tag{2}$$

Cohen and Migliorati [11] prove that under suitable conditions the RIP (2) is satisfied with high probability.

Theorem 1.1. *If* $K_{\boldsymbol{B},w} := \operatorname{ess\,sup}_{y\in Y} w(y)\boldsymbol{B}(y)^\intercal\boldsymbol{B}(y) < \infty$ *then*

$$\mathbb{P}[\|\mathbf{G}-\mathbf{I}_m\|_2 > \delta] \leq 2m\exp\left(-\frac{c_\delta n}{K_{\boldsymbol{B},w}}\right)$$

with $c_\delta := \delta + (1-\delta)\ln(1-\delta)$.

The notion of a RIP was introduced in the context of compressed sensing [4]. It expresses the well-posedness of the problem by ensuring that $\|\bullet\|_n$ is indeed a norm and thus a convergence of the empirical norm implies a convergence in the real norm. In compressed sensing of sparse vectors [4, 5] and low-rank tensors [7] discrete analogues of the RIP (2) are employed to derive bounds for the corresponding reconstruction errors. A recent work which generalizes the RIP from [11] to sparse grid spaces is [12].

In this paper we extend the cited results to more general norms and nonlinear model sets by directly bounding the probability of

$$\mathrm{RIP}_A(\delta) :\Leftrightarrow (1-\delta)\|u\|^2 \leq \|u\|_n^2 \leq (1+\delta)\|u\|^2 \qquad \forall u \in A.$$

We prove that this RIP holds with high probability and use it to provide quasi-optimality guarantees in the considered function approximation setting.

In Remark 2.5 we note that $\mathrm{RIP}_A(\delta) \Leftrightarrow \mathrm{RIP}_{\mathrm{Cone}(A)}(\delta)$. This means that it suffices to consider conic model sets. Optimizing over these sets is not straight-forward and [13] derive sufficient RIP constants for exact recovery of conic model sets using a suitable regularizer.

2. Main Result

To measure the rate of convergence with which $\|v\|_n$ approaches $\|v\|$ as n tends to ∞, we introduce the *variation constant*

$$K(A) := \sup_{u\in A}\|u\|_{w,\infty}^2 \qquad \text{with} \qquad \|v\|_{w,\infty}^2 := \operatorname*{ess\,sup}_{y\in Y} w(y)|v|_y^2.$$

This constant constitutes a uniform upper bound of $\|v\|_n$ for all realizations of the empirical norm $\|\bullet\|_n$ and all $v \in A$. We usually omit the dependence of K on the choice of w, $|\bullet|_y$ and Y. When a distinction between different choices of these parameters is necessary we add suitable subscripts to K.

Remark 2.1. *The bias-variance trade-off is directly reflected in the definition of* K*. Enlarging the model set* A *reduces the approximation error but at the same time increases the variation constant* K *and thereby decreases the rate of convergence of the empirical norm on* A*.*

The constant K is a fundamental parameter in many concentration inequalities that are used to provide bounds for the rate of convergence of the *quadrature error*.

Definition 2.2 (Quadrature Error)**.** *The* quadrature error *of the empirical norm $\|\bullet\|_n^2$ on the model set $A \subseteq \mathcal{V}$ is defined by*

$$\mathcal{E}_A := \sup_{u\in A} |\|u\|^2 - \|u\|_n^2|.$$

This error is closely related to the RIP. In order to see this relation, we introduce a *normalization operator* U.

Definition 2.3 (Normalization)**.** *The* normalization operator *acts on a set A by*

$$U(A) := \operatorname{Cone}(A) \cap \mathrm{S}_1^{\mathcal{V}}(0) = \left\{ \tfrac{u}{\|u\|} : u \in A \setminus \{0\} \right\},$$

where $\operatorname{Cone}(A) := \{\alpha a : \alpha \in (0,\infty) \wedge a \in A\}$ denotes the cone generated by A and $\mathrm{S}_r^{\mathcal{V}}(c) := \{v \in \mathcal{V} : \|x - c\| = r\}$ denotes the sphere of radius r and centre c in $\mathcal{V}$.

With this definition the variation constant $K(U(A))$ can be seen as a generalization of the embedding constant $(A, \|\bullet\|) \hookrightarrow (A, \|\bullet\|_{w,\infty})$ to nonlinear sets and therefore as an analog of $K_{\boldsymbol{B},w}$ in Theorem 1.1. Using the normalization operator the subsequent lemma follows almost immediately.

Lemma 2.4 (Equivalence of RIP and generalization error bound)**.** *For some set A,*

$$\mathrm{RIP}_A(\delta) \Leftrightarrow \mathcal{E}_{U(A)} \leq \delta \quad \text{for } \delta > 0.$$

Proof. Note that $\|\alpha u\|_n = |\alpha| \|u\|_n$ for all $u \in A$ and $\|u\| = 1$ for all $u \in U(A)$. Therefore,

$$\begin{aligned}
(1-\delta)\|u\|^2 &\leq \|u\|_n^2 &&\leq (1+\delta)\|u\|^2 && \forall u \in A \\
\Leftrightarrow \quad (1-\delta) &\leq \left\|\tfrac{u}{\|u\|}\right\|_n^2 &&\leq (1+\delta) && \forall u \in A \\
\Leftrightarrow \quad -\delta &\leq \|u\|_n^2 - \|u\|^2 &&\leq \delta && \forall u \in U(A),
\end{aligned}$$

which holds exactly if $\sup_{u\in U(A)} |\|u\|^2 - \|u\|_n^2| \leq \delta$. ∎

Remark 2.5. *Lemma 2.4 implies that*

$$\mathrm{RIP}_A(\delta) \Leftrightarrow \mathrm{RIP}_{\operatorname{Cone}(A)}(\delta) \quad \text{for } \delta > 0,$$

and consequently also holds for unbounded A.

We introduce the notion of a covering number to provide a well-known bound for the quadrature error in the following.

Definition 2.6 (Covering Number)**.** *The covering number $\nu_{\|\bullet\|}(A,\varepsilon)$ of a subset $A \subseteq \mathcal{V}$ is the minimal number of $\|\bullet\|$-open balls of radius ε needed to cover A.*

Lemma 2.7. *Let $A \subseteq \mathcal{V}$ and $|\bullet|_y$ be such that $K = K(U(A)) < \infty$. Then,*

$$\mathbb{P}[\mathcal{E}_{U(A)} \geq \delta] \leq 2\nu_{\|\bullet\|_{w,\infty}}\left(U(A), \tfrac{1}{8}\tfrac{\delta}{\sqrt{K}}\right) \exp\left(-\tfrac{n}{2}\left(\tfrac{\delta}{K}\right)^2\right) \quad \text{for } \delta > 0.$$

The proof of this lemma can be found in Appendix A. With the preceding preparations we can derive a central result:

Theorem 2.8. *Let $A \subseteq \mathcal{V}$ and $|\bullet|_y$ be such that $K = K(U(A)) < \infty$. Then,*

$$\mathbb{P}[\mathrm{RIP}_A(\delta)] \geq 1 - 2\nu_{\|\bullet\|_{w,\infty}}\Big(U(A), \tfrac{1}{4}\tfrac{\delta}{\sqrt{K}}\Big)\exp\Big(-\tfrac{n}{2}\Big(\tfrac{\delta}{K}\Big)^2\Big) \quad \text{for } \delta > 0.$$

Proof. By Lemma 2.4 it suffices to bound the quadrature error on $U(A)$. Lemma 2.7 provides a bound for the probability of the converse event. ∎

Corollary 2.9 (Sample Complexity)**.** *Let $M > 0$ and $A \subseteq \mathcal{V}$ be a set with $\nu_{\|\bullet\|_{w,\infty}}(U(A), r) \leq \frac{1}{2}r^{-M}$. Let p be defined as in Theorem 2.8. Under the assumptions of Theorem 2.8 with $K = K(U(A))$, at least*

$$n = 2\left(M\ln\left(4\frac{\sqrt{K}}{\delta}\right) - \ln(p)\right)\left(\frac{K}{\delta}\right)^2$$

many samples are required to satisfy $\mathrm{RIP}_A(\delta)$ with a probability larger than $1 - p$.

Proof. To obtain $\mathrm{RIP}_{U(A)}(\delta)$ with a probability of at least $1 - p$ it suffices that

$$\begin{aligned} p &= 2\nu(U(A), \tfrac{1}{4}\tfrac{\delta}{\sqrt{K}})\exp(-\tfrac{n}{2}\Big(\tfrac{\delta}{K}\Big)^2) \\ &= \exp\Big(M\ln\Big(4\tfrac{\sqrt{K}}{\delta}\Big) - \tfrac{n}{2}\Big(\tfrac{\delta}{K}\Big)^2\Big). \end{aligned}$$

Equivalently,

$$\begin{aligned} &\ln p = M\ln\Big(4\tfrac{\sqrt{K}}{\delta}\Big) - \tfrac{n}{2}\Big(\tfrac{\delta}{K}\Big)^2 \\ \Leftrightarrow \qquad &n = 2(M\ln\Big(4\tfrac{\sqrt{K}}{\delta}\Big) - \ln p)\Big(\tfrac{K}{\delta}\Big)^2. \end{aligned}$$

∎

Linear spaces, sparse vectors and low-rank tensors all satisfy the requirements of this corollary with M depending linearly on the number of parameters of the model [7, 9, 14]. The corollary states that in these cases $n \in \mathcal{O}(MG)$ depends linearly on the number of parameters M with a factor $G := \ln(K)K^2$ representing the variation of $\|\bullet\|_n$ on $\mathcal{M}$.

Remark 2.10. *Corollary 2.9 shows that the variation constant K is of greater importance than the covering number ν which enters the bound on the sample complexity only logarithmically.*

Example 2.11 (K represents regularity and not dimension)**.** *One might think that the convergence of $\|\bullet\|_n \to \|\bullet\|$ should only depend on the interior dimension of the model set $\mathcal{M}$ and not on the dimension of the ambient space $\mathcal{V}$. However, a counter-example can be constructed easily. A sphere-filling rope is a closed differentiable curve $\gamma : \mathrm{S}_1^{\mathbb{R}^2}(0) \to \mathrm{S}_1^{\mathbb{R}^3}(0)$ together with a radius $r > 0$ such that at every point γ has a distance of $2r$ to itself. An illustration of this is provided in Figure 1 and a classification of such curves can be found in [15]. The set $A := \gamma(\mathrm{S}_1^{\mathbb{R}^2}(0))$ is a smooth manifold of dimension 1 but $K(U(A))$ approaches $K(U(\mathcal{V}))$ when the radius r goes to zero.*

We derive an estimate of the error due to the empirical evaluation of the projection which can be obtained when a RIP is satisfied. Note that for the sake of a simpler notation we henceforth use $u + \mathcal{M} := \{u\} + \mathcal{M}$ with a single element u.

Theorem 2.12 (Empirical Projection Error)**.** *Assume that $\mathrm{RIP}_{Pu-\mathcal{M}}(\delta)$ holds. Then*

$$\|(P - P_n)u\| \leq 2\frac{1}{\sqrt{1-\delta}}\|(1-P)u\|_{w,\infty}.$$

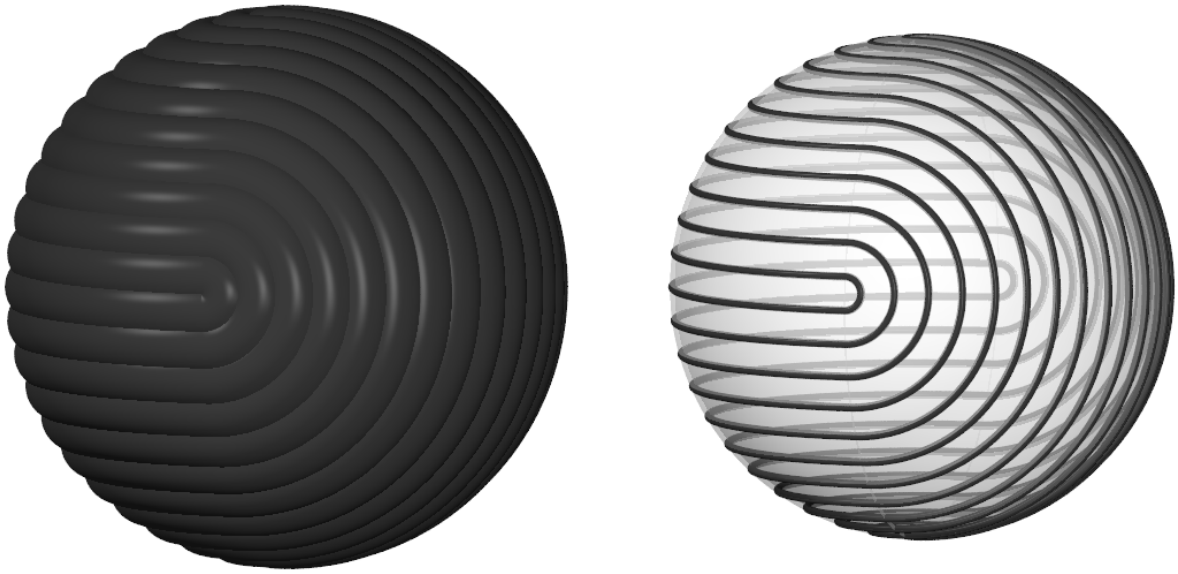


Figure 1. A sphere-filling rope (left) and its corresponding curve (right).

If in addition $\mathrm{RIP}_{\{(1-P)u\}}(\delta)$ *is satisfied then*

$$\|(P-P_n)u\| \leq 2\sqrt{\frac{1+\delta}{1-\delta}}\|(1-P)u\|.$$

Proof. First observe that $P_n u \in \mathcal{M}$ and therefore $(P-P_n)u \in Pu-\mathcal{M}$. Consequently,

$$\begin{aligned}\|(P-P_n)u\| &\leq \frac{1}{\sqrt{1-\delta}}\|(P-P_n)u\|_n \\ &\leq \frac{1}{\sqrt{1-\delta}}[\|(P-1)u\|_n + \|(1-P_n)u\|_n] \\ &\leq 2\frac{1}{\sqrt{1-\delta}}\|(1-P)u\|_n,\end{aligned}$$

where each inequality follows from $\mathrm{RIP}_{Pu-\mathcal{M}}(\delta)$, the triangle inequality and the definition of P_n, respectively.

The first assertion holds since $\|v\|_n \leq \|v\|_{n,\infty}$ is satisfied for all $v \in \mathcal{V}$ and in particular for $(1-P)u$. The second assertion follows by an application of $\mathrm{RIP}_{\{(1-P)u\}}(\delta)$. ■

Remark 2.13. *If* $u \in \mathcal{M}$ *then* $((1-P)u) \in -(Pu-\mathcal{M})$ *and* $\mathrm{RIP}_{Pu-\mathcal{M}}(\delta)$ *implies* $\mathrm{RIP}_{\{(1-P)u\}}(\delta)$.

Remark 2.14 (Reconstruction with Noise)**.** *Consider the randomly perturbed seminorm* $|v|_y + \eta_y$ *where* η_y *is a centered random process satisfying the bound* $w(y)\eta_y^2 \leq \frac{1}{4}(1-\delta)\varepsilon^2$ *for some* $\varepsilon > 0$ *and* $\delta \in (0,1)$*. This seminorm induces the perturbed empirical norm*

$$\|v\|_{\eta,n} := \left(\frac{1}{n}\sum_{i=1}^n w(y_i)(|v|_{y_i} + \eta_{y_i})^2\right)^{1/2}$$

and the perturbed empirical best approximation operator

$$P_{\eta,n}u \in \operatorname*{arg\,min}_{v\in\mathcal{M}}\|u-v\|_{\eta,n}.$$

Assume that $\mathrm{RIP}_{Pu-\mathcal{M}}(\delta)$ *holds. Then*

$$\|(P - P_{\eta,n})u\| \leq 2\frac{1}{\sqrt{1-\delta}}\|(1-P)u\|_{w,\infty} + \varepsilon.$$

If in addition $\mathrm{RIP}_{\{(1-P)u\}}(\delta)$ *is satisfied then*

$$\|(P - P_{\eta,n})u\| \leq 2\sqrt{\frac{1+\delta}{1-\delta}}\|(1-P)u\| + \varepsilon.$$

Note that the projection error can be split into an approximation and an estimation error by the triangle inequality. It immediately follows that

$$\begin{aligned}\|(1-P_n)u\| &\leq \|(1-P)u\| + \|(P-P_n)u\| \\ &\leq (1 + 2\frac{\sqrt{1+\delta}}{\sqrt{1-\delta}})\|(1-P)u\|.\end{aligned}$$

Hence, under suitable assumptions the empirical projection is quasi optimal. Depending on the considered problem, the best approximation error $\|(1-P)u\|$ is usually covered by results in functional and numerical analysis.

Remark 2.15 (Deterministic Samples)**.** *Theorem 2.12 is also valid for deterministic instead of random samples, e.g. determined by some quadrature formula. In this case, it has to be verified that the chosen sample set satisfies the RIP.*

Remark 2.16 (Error Equilibration)**.** *In an adaptive scheme the estimation error* $\|(P-P_n)u\|$ *only has to be minimized to the same extent as the approximation error* $\|(1-P)u\|$ *in order to equilibrate error contributions. Corollary 2.9 implies that it suffices to use a fixed* $\delta < 1$ *and raise the number of samples* n *only to increase the probability of the RIP. In [11] the empirical Gramian could be used to verify this RIP for a given sample set. In the nonlinear setting this is no longer possible. To obtain an indicator for the convergence of our method we make the following considerations. Define* $A := (Pu - \mathcal{M}) \cup \{(1-P)u\}$, $e_n := \|(1-P_n)u\|$ *and* $e := \|(1-P)u\|$*. Observe that for* $\delta \leq \frac{1}{\sqrt{2}}$

$$1 + \delta \leq \sqrt{\frac{1+\delta}{1-\delta}} \leq 1 + 2\delta.$$

Combining the second inequality with Theorem 2.12 leads to

$$\begin{aligned}\mathrm{RIP}_A(\delta) \Rightarrow e_n &\leq \left(1 + 2\sqrt{\tfrac{1+\delta}{1-\delta}}\right)e \leq (1 + 2(1+2\delta))e \\ &\Rightarrow e_n \leq (3+4\delta)e.\end{aligned}$$

Therefore,

$$\mathbb{P}[\mathrm{RIP}_A(\delta)] \leq \mathbb{P}[e_n \leq (3+4\delta)e]. \tag{3}$$

By Theorem 2.8 there exist c *and* $\nu(\delta)$ *such that*

$$1 - \nu(\delta)\exp(-cn\delta^2) \leq \mathbb{P}[\mathrm{RIP}_A(\delta)].$$

Combining this with (3) *yields*

$$1 - \nu(\delta)\exp(-cn\delta^2) \leq \mathbb{P}[e_n \leq (3+4\delta)e] =: p(\delta).$$

Since $p(\delta)$ *is increasing in* δ *we can define an inverse* $\delta(p)$ *in the sense of the quantile function and rearrange the last equation as*

$$-\ln(1-p) \geq cn\delta(p)^2 - \ln(\nu(\delta(p))).$$

Both $\delta(p)$ and $-\ln(\nu(\delta(p))) \leq 0$ are increasing. This means that for large p the second term in the above sum becomes negligible, yielding

$$\delta(p) \lesssim n^{-1/2}$$

and consequently

$$e_n \lesssim (1 + n^{-1/2})e.$$

Thus, in this regime, we obtain the classical Monte Carlo convergence rate of $\mathcal{O}(n^{-1/2})$. Conversely, we can use this rate as an indicator for equivalence of the estimation error and the approximation error. When $\mathcal{O}(n^{-1/2})$-convergence is observed, equivalence is assumed and additional sampling can be deemed unnecessary. This behaviour is illustrated in Figure 2.

This also highlights that the model class only influences the rate with which the RIP is satisfied. As soon as the RIP holds, any further convergence of the error only achieves the slow Monte Carlo rate.

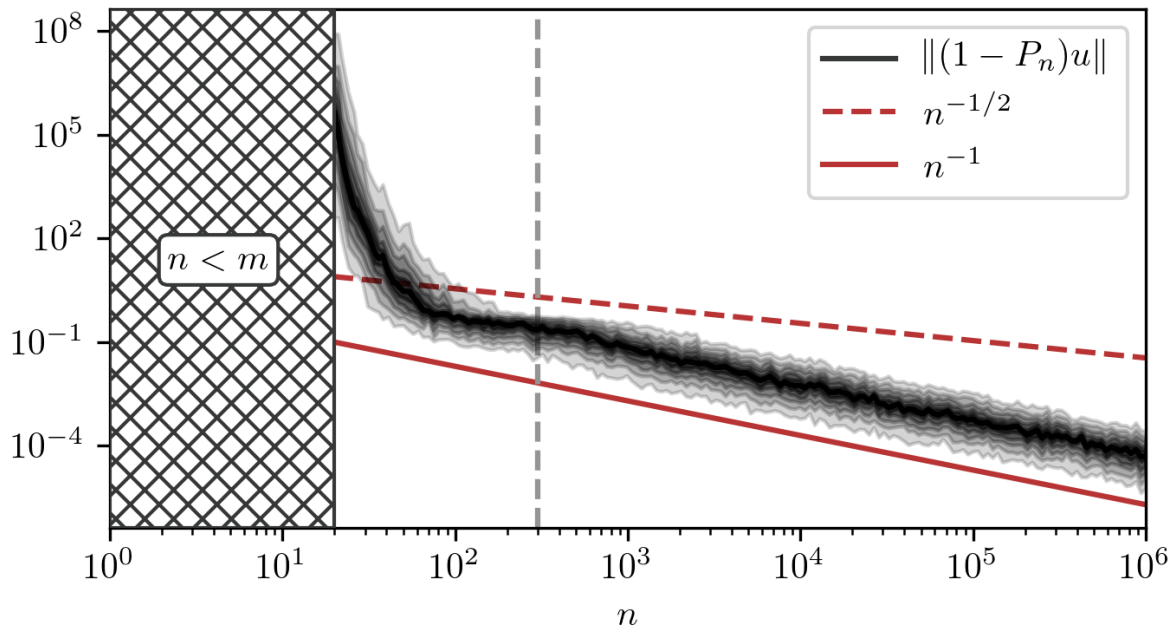


Figure 2. Depicted is the error of an empirical approximation in the model space m of polynomials of degree less than 20 (i.e. $M = 20$) in relation to n. The hatched area on the left marks a range of n where the approximation problem is underdetermined and any error can be reached. When $n \geq m$ the approximation problem has a unique solution in the least squares sense. From this point until the gray and dashed line an exponential decay of the error can be observed. This decay results from the exponentially fast convergence of the probability for the RIP w.r.t. n. From there on the RIP holds with a high probability and the error decays with a rate of n^{-1}. This faster decay can be explained as a pre-asymptotic phenomenon by using Bernstein's inequality instead of Hoeffding's inequality in the proof of Theorem 2.8.

3. Examples and numerical illustrations

In this section, we examine some exemplary model spaces to which the developed theory can be applied. More specifically, we consider linear spaces, sparse vectors and tensors of fixed rank. The following theorem is central to the further considerations.

Theorem 3.1. *Let the ambient vector space $\mathcal{V}$ be separable and $A \subseteq \mathcal{V}$. Then the pointwise supremum $\hat{b}(y) \coloneqq \sup_{v\in A}|v|_y^2$ with respect to $y \in Y$ is measurable and*

$$K(A) = \|w\hat{b}\|_{L^\infty(Y,\rho)}.$$

With this, an optimal (a priori) weight function is given by $w = \|\hat{b}\|_{L^1(Y,\rho)}\hat{b}^{-1}$.

Proof. See Appendix B. ∎

This Theorem allows to analyse the seminorm and the model class independently from the choice of weight function which can be chosen optimally when these first two parameters are fixed.

3.1. **Linear Operators and Energy Spaces.** Let $\mathcal{H} \subseteq \mathcal{V}$ be a subspace on which the seminorm takes the form $|u|_y := \|L_y u\|_2$ for $m \in \mathbb{N}$ and a family of bounded y-dependent linear operators $L_y \in \mathcal{L}(\mathcal{H}, \mathbb{R}^m)$. For elliptic differential operators $A = L_y^T L_y$, e.g. with $L_y := \nabla$ and the Laplacian $\Delta = L^T L$, this results in a corresponding Dirichlet energy. Then, for all $u \in \mathcal{H}$,

$$|u|_y^2 = \|L_y u\|_2^2 \leq \|L_y\|^2_{\mathcal{L}(\mathcal{H},\mathbb{R}^m)}\|u\|^2_{\mathcal{H}}.$$

This is a generalization of the concept of a *reproducing kernel Hilbert space (RKHS)* and implies that for any $A \subseteq \mathcal{H}$

$$\hat{b}(y) = \kappa(y)K^{\mathcal{H}}(U(A)),$$

with $\kappa(y) := \|L_y\|^2_{\mathcal{L}(\mathcal{H},\mathbb{R}^m)}$ and $K^{\mathcal{H}}(U(A)) := \sup_{u\in A}\frac{\|u\|^2_{\mathcal{H}}}{\|u\|^2}$. This decouples the choice of the seminorm represented by the factor κ and the choice of the model class represented by the factor $K^{\mathcal{H}}$.

Remark 3.2. *In this setting the application of Theorem 2.12 leads to*

$$\|(1-P_n)u\| \lesssim \|(1-P)u\|_{w,\infty} \leq \|w\kappa\|_{L^\infty(Y)}\|(1-P)u\|_{\mathcal{H}}$$

whenever $\mathrm{RIP}_{Pu+m}(\delta)$ holds.

Example 3.3. *Let $\mathcal{V}_m \subseteq \mathcal{V} := L^2(Y,\rho)$ be the m-dimensional linear subspace spanned by the continuous* orthonormal *basis $\{\mathbf{B}_j\}_{j\in[m]}$. This is a RKHS with reproducing kernel $k(x,y) := \boldsymbol{B}^\intercal(x)\boldsymbol{B}(y)$. It is well known that $\kappa(y) = k(y,y)$ and thus*

$$K(U(\mathcal{V}_m)) = \operatorname*{ess\,sup}_{y\in\mathbb{R}^d} w(y)\boldsymbol{B}^\intercal(y)\boldsymbol{B}(y) = K_{\boldsymbol{B},w}.$$

According to Theorem 3.1, the optimal choice $w(y) \coloneqq m(\boldsymbol{B}^\intercal(y)\boldsymbol{B}(y))^{-1}$ leads to $K(U(\mathcal{V}_m)) = m$. This was also observed in [11].

Using the fact that $\|v\|_{w,\infty} \leq \sqrt{K}\|v\|$ and therefore

$$\nu_{\|\bullet\|_{w,\infty}}(U(\mathcal{V}_m), r) \leq \nu_{\|\bullet\|}\left(U(\mathcal{V}_m), \frac{r}{\sqrt{K}}\right) \leq \left(\frac{2\sqrt{mK}}{r}\right)^m,$$

we can bound the sample complexity of this model class by Corollary 2.9. This bound is similar to (but slightly weaker than) the bound provided in [11].

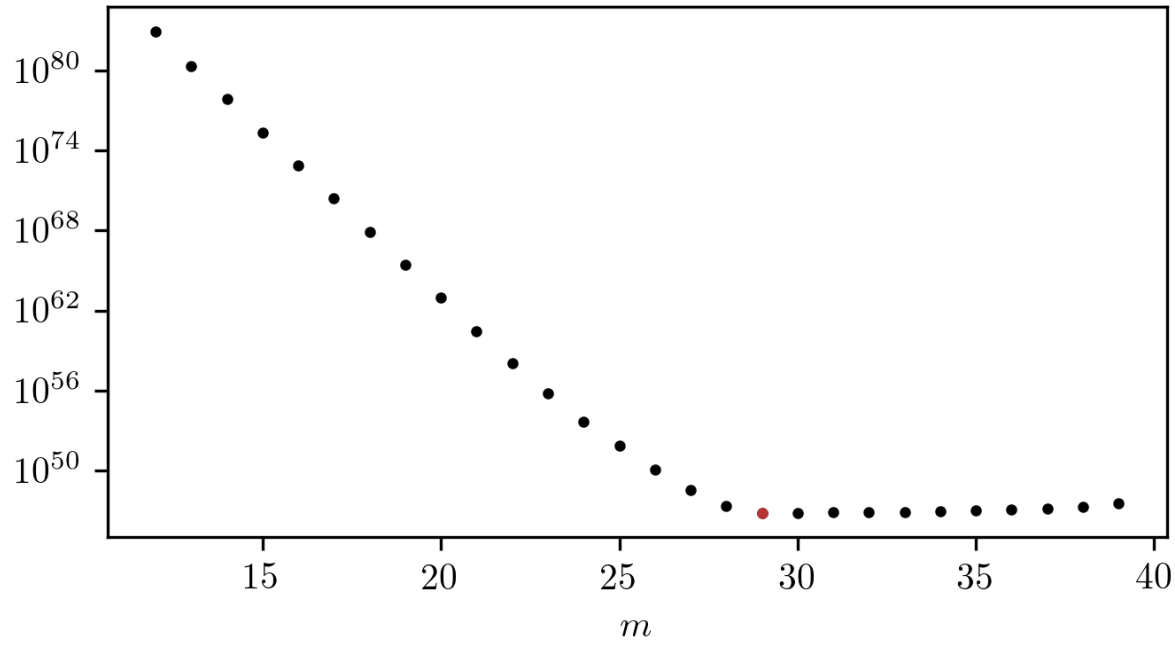


Figure 3. The upper bound $\kappa_m \lambda_m$ for the variation constant. The optimal m is marked in red.

Example 3.4. *Spaces with higher regularity can be endowed with stronger norms which may lead to a lower sample complexity. To see this let, Y be a Lipschitz domain in $\mathbb{R}^d$ and $\mathcal{V} := H^m(Y)$. Assume $M \geq m + \frac{d}{2}$ and $A \subseteq \mathcal{H} := H^M(Y)$. When m increases the seminorm captures more of the regularity of $\mathcal{H}$ and the constant*

$$\lambda_m := K^{\mathcal{H}}(U(A)) = \sup_{u \in A} \frac{\|u\|^2_{H^M(Y)}}{\|u\|^2_{H^m(Y)}}$$

decreases. Increasing m however also increases the complexity of the seminorm and consequently the value of

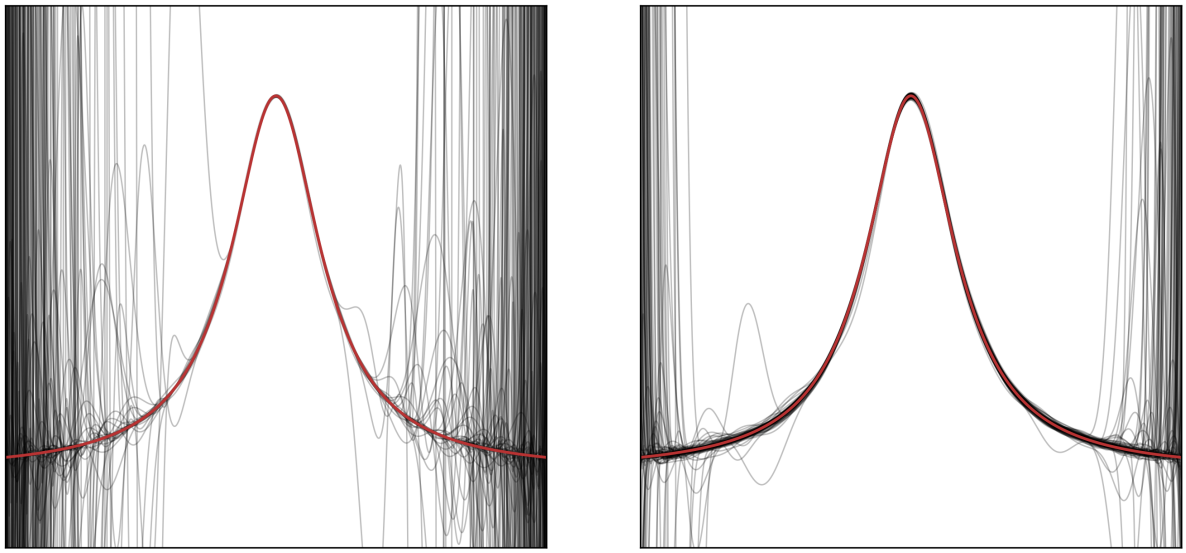
$$\kappa(y) \leq \varkappa_m := (2\sqrt{\pi})^{-d} \frac{\Gamma(M+1)\Gamma(M-m-\frac{d}{2})}{\Gamma(M-m)}.$$

A proof of this inequality is provided in Appendix C. In practice we therefore need to find a value of m where both effects are equilibrated. This is illustrated in Figure 3. We conclude that an approximation with respect to the H^1-norm converges faster than an approximation with respect to the L^2-norm which is illustrated numerically in Figure 4.

Note also that the minimization with respect to the $H^k(Y,\rho)$-norm does not necessarily require more computational effort than the minimization with respect to the $L^2(Y,\rho)$-norm. The values of both seminorms can be computed with a single evaluation of the Fourier transform $\mathcal{F}u$ of u.

Remark 3.5. *For $\mathcal{V} = L^2([-1,1], \frac{\mathrm{d}x}{2}; \mathbb{C})$ consider the Fourier basis. This is an orthonormal basis for which the basis functions are bounded by 1 almost everywhere. This implies that the variation constant in $\mathcal{V}$ is 1. Other bases that satisfy these simultaneous orthogonality and boundedness conditions are investigated in [16].*

3.2. **Sets of smooth functions.** In this section we provide a bound for functions with high regularity following the ideas in [17]. For $Y = [0,1]^d$ and $\mathcal{V} := L^2(Y, \mathrm{d}y)$, we consider the class of μ-smooth



(a) Least squares, L^2 **(b)** Least squares, H^1

Figure 4. Overlaid least squares approximations of the function $f(x) = \frac{1}{1+25x^2}$ (red) by Legendre polynomials of degree 29. Different approximations correspond to different random draws of $n = 40$ sampling points from the uniform measure on $[-1, 1]$.

functions

$$\mathcal{S}_\mu := \{v \in \mathcal{V} : \lim_{m\to\infty} |v|_m^{1/m} \leq \mu\}, \qquad \mu > 0,$$

where the H^m-seminorm is defined as

$$|v|_m := \sqrt{\sum_{k=1}^{d} \left\| \frac{\partial^m v}{\partial y_k^m} \right\|^2}.$$

In the following we assume $w \equiv 1$. This implies $\|v\|_{w,\infty} = \|v\|_{L^\infty(Y,\mathrm{d}y)}$.

Lemma 3.6. *It holds that*

- $\mathcal{S}_\mu \subseteq \mathcal{S}_\nu$ *for* $\mu \leq \nu$,
- $\mathcal{S}_\mu = -\mathcal{S}_\mu$,
- $\mathcal{S}_\mu + \mathcal{S}_\nu \subseteq \mathcal{S}_{\max\{\mu,\nu\}}$ *and*
- $\|v\|_{w,\infty} \leq c^Y \max\{1, \mu^{d/2}\}\|v\|$ *for all* $v \in \mathcal{S}_\mu$.

Proof. The first two properties are trivial. To prove the third, consider $v_\mu \in \mathcal{S}_\mu$ and $v_\nu \in \mathcal{S}_\nu$ and observe that

$$|v_\mu + v_\nu|_m \leq |v_\mu|_m + |v_\nu|_m \leq 2\max\{|v_\mu|_m, |v_\nu|_m\}.$$

Consequently,

$$\begin{aligned}\lim_{m\to\infty} |v_\mu + v_\nu|_m^{1/m} &\leq \left(\lim_{m\to\infty} 2^{1/m}\right) \cdot \left(\limsup_{m\to\infty} \max\{|v_\mu|_m^{1/m}, |v_\nu|_m^{1/m}\}\right) \\ &\leq \max\{\limsup_{m\to\infty} |v_\mu|_m^{1/m}, \limsup_{m\to\infty} |v_\nu|_m^{1/m}\} \\ &\leq \max\{\mu, \nu\}.\end{aligned}$$

To prove the last property we consider $v \in \mathcal{S}_\mu$ and use the Gagliardo-Nirenberg interpolation inequality (cf. [17])

$$\|v\|_{w,\infty} \leq c_m^Y \Big[|v|_m^2 + \|v\|^2\Big]^{d/4m} \|v\|^{1-d/2m}.$$

Taking the limit $m \to \infty$ results in the estimate

$$\begin{aligned}\|v\|_{w,\infty} &\leq \left(\lim_{m\to\infty} c_m^Y\right) \cdot \left(\lim_{m\to\infty} [|v|_m^2 + \|v\|^2]^{d/4m}\right) \cdot \left(\lim_{m\to\infty} \|v\|^{1-d/2m}\right) \\ &\leq c^Y \cdot \left(\lim_{m\to\infty} \Big(2\max\{|v|_m^2, \|v\|^2\}\Big)^{d/4m}\right) \cdot \|v\| \\ &\leq c^Y \cdot \max\Big\{1, \mu^{d/2}\Big\} \cdot \|v\|,\end{aligned}$$

where c^Y is defined as $c^Y := \lim_{m\to\infty} c_m^Y$. ■

Lemma 3.7. $K(U(\mathcal{S}_\mu)) \leq \left(c^Y \max\{1, \mu^{d/2}\}\right)^2$.

Proof. This follows directly from Lemma 3.6. ■

3.3. **Sets of sparse functions.** In this section we follow the ideas of [6] and consider spaces with weighted sparsity constraints. For any sequence $\omega \in \mathbb{R}^{\mathbb{N}}_{\geq 0}$ and any subset $S \subseteq \mathbb{N}$, define a weighted cardinality and a weighted ℓ^0-seminorm by

$$\omega(S) := \sum_{j\in S} \omega_j^2 \qquad \text{and} \qquad \|\mathbf{v}\|_{\omega,0} := \omega(\operatorname{supp}(\mathbf{v})).$$

Observe that $\omega_j \preceq \tilde{\omega}_j$ implies $\omega(S) \leq \tilde{\omega}(S)$ and that $\omega(S) = |S|$ for $\omega \equiv \mathbf{1}$. This warrants the names cardinality and ℓ^0-seminorm.

Let $\{\mathbf{B}_j\}_{j\in\mathbb{N}}$ be a fixed *orthonormal* basis for $\mathcal{V} := L^2(Y, \rho)$ and fix a weight function w. For any sequence ω with $\omega_j \geq \|\mathbf{B}_j\|_{w,\infty}$ we define the model set

$$\mathcal{M}_{\omega,s} := \{v \in \mathcal{V} : \|\mathbf{v}\|_{\omega,0} \leq s\},$$

where $\mathbf{v}$ denotes the coefficient vector of $v \in \mathcal{V}$ with respect to the basis $\{\mathbf{B}_j\}_{j\in\mathbb{N}}$.

Lemma 3.8. *It holds that*

- $\mathcal{M}_{\omega,s} \subseteq \mathcal{M}_{\omega,t}$ *for* $s \leq t$,
- $\mathcal{M}_{\omega,s} = -\mathcal{M}_{\omega,s}$,
- $\mathcal{M}_{\omega,s} + \mathcal{M}_{\omega,t} \subseteq \mathcal{M}_{\omega,s+t}$ *and*
- $\|v\|_{w,\infty} \leq \sqrt{s}\|v\|$ *for all* $v \in \mathcal{M}_{\omega,s}$.

Proof. The first three assertions are trivial. To prove the last one let $v \in \mathcal{M}_{\omega,s}$. Using the triangle inequality and $\omega_j \geq \|\mathbf{B}_j\|_{w,\infty}$, we obtain

$$\|v\|_{w,\infty} \leq \sum_{j=1}^{\infty} |\mathbf{v}_j| \|\mathbf{B}_j(y)\|_{w,\infty} \leq \sum_{j=1}^{\infty} |\mathbf{v}_j|\omega_j = \sum_{j\in\operatorname{supp}(\mathbf{v})} |\mathbf{v}_j|\omega_j.$$

The Cauchy-Schwarz inequality, $\|\mathbf{v}\|_{\omega,0} \leq s$ and the orthonormality of $\mathbf{B}$ yield

$$\leq \|\mathbf{v}\|_2 \sqrt{\sum_{j\in\operatorname{supp}(\mathbf{v})} \omega_j^2} = \|\mathbf{v}\|_2 \sqrt{\|\mathbf{v}\|_{\omega,0}} \leq \|v\|\sqrt{s}.$$

■

Lemma 3.9. $K(U(\mathcal{M}_{\omega,s})) \leq s$.

Proof. This follows directly from Lemma 3.8. ■

Remark 3.10. *The constant sequence* $\omega \equiv \omega_{\max} := \max_{j\in[m]} \|\mathbf{B}_j\|_{w,\infty}$ *represents the set* $\mathcal{M}_{\omega,s} = \mathcal{M}_{\mathbf{1},\lfloor s/\omega_{\max}\rfloor}$ *of* $\lfloor \frac{s}{\omega_{\max}} \rfloor$*-sparse functions. When the chosen basis is a tensor product basis* $\mathbf{B}_\mathbf{j} := B_{j_1} \otimes \cdots \otimes B_{j_d}$ *this means that*

$$s \geq \max_{\mathbf{j}\in[m]^d} \|\mathbf{B}_\mathbf{j}\|_{w,\infty} = \left(\max_{j\in[m]} \|B_j\|_{w,\infty}\right)^d$$

grows exponentially with the dimension. This is a drawback of the standard definition of sparsity.

Lemma 3.11. *Let* $\mathcal{V}_m$ *be an* m*-dimensional subspace spanned by a subset of* $\{\mathbf{B}_j\}_{j\in\mathbb{N}}$*. Then*

$$\nu_{\|\bullet\|_{w,\infty}}(U(\mathcal{M}_{\omega,s} \cap \mathcal{V}_m), r) \lesssim \left(\tfrac{m}{r\sqrt{s}}\right)^s.$$

Proof. We show that

$$\nu_{\|\bullet\|_{w,\infty}}(U(\mathcal{M}_{\omega,s} \cap \mathcal{V}_m), r) \leq \nu_{\|\bullet\|}\left(U(\mathcal{M}_{\omega,s} \cap \mathcal{V}_m), \frac{r}{\sqrt{2s}}\right) \lesssim \left(\tfrac{m}{r\sqrt{s}}\right)^s.$$

For the first step, let $\{v_j\}$ be the centers of a $\|\bullet\|$-covering of $U(\mathcal{M}_{\omega,s} \cap \mathcal{V}_m)$ with radius $\frac{r}{\sqrt{2s}}$. Thus, for any $v \in U(\mathcal{M}_{\omega,s} \cap \mathcal{V}_m)$ there exists v_j such that $\|v - v_j\| \leq \frac{r}{\sqrt{2s}}$. Since $v - v_j \in \mathcal{M}_{\omega,2s}$ and by Lemma 3.8,

$$\|v - v_j\|_{w,\infty} \leq \sqrt{2s}\|v - v_j\|_{L^2} \leq r.$$

This implies that $\{v_j\}$ are also the centers of an $\|\bullet\|_{w,\infty}$-covering with radius r.

For the second step, observe that $\mathcal{M}_{\omega,s} \subseteq \mathcal{M}_{\mathbf{1},s} = \mathcal{M}_{\mathbf{1},\lfloor s\rfloor}$. Since $(\mathcal{V}_m, \|\bullet\|) \simeq (\mathbb{R}^m, \|\bullet\|_2)$ it remains to compute the covering number for the unit sphere of $\lfloor s\rfloor$-sparse vectors in $\mathbb{R}^m$. A bound for this is given in [14] by

$$\nu_{\|\bullet\|_2}\left(S_1^{\mathbb{R}^m}(0) \cap \mathcal{M}_{\mathbf{1},\lfloor s\rfloor}, \frac{r}{\sqrt{2s}}\right) \lesssim \left(\frac{d\sqrt{s}}{r\lfloor s\rfloor}\right)^{\lfloor s\rfloor} \leq \left(\frac{d}{r\sqrt{s}}\right)^s.$$

■

Theorem 3.12. *Let* $\mathcal{V}_m$ *be an* m*-dimensional subspace spanned by a subset of* $\{\mathbf{B}_j\}_{j\in\mathbb{N}}$*. Then,*

$$\mathbb{P}[\neg\,\mathrm{RIP}_{\mathcal{M}_{\omega,s}\cap\mathcal{V}_m}(\delta)] \lesssim \exp\left(s\ln\left(\tfrac{m}{\delta}\right) - \tfrac{n}{2}\left(\tfrac{\delta}{s}\right)^2\right).$$

Proof. The assertion follows directly from Theorem 2.8 together with Lemmas 3.9 and 3.11. ■

If the sequence ω is increasing, the intersection $\mathcal{M}_{\omega,s} \cap \mathcal{V}_m$ required in Theorem 3.12 occurs naturally since $\mathcal{M}_{\omega,s}$ must be contained in the finite-dimensional linear space

$$\mathcal{V}_m := \operatorname{span}\{\mathbf{B}_j : \omega_j^2 \leq s\}.$$

In this way the choice of ω influences the sample complexity by controlling the growth of the dimension of $\mathcal{V}_m$. Using the basis of Legendre polynomials and $\omega_j := \|\mathbf{B}_j\|_{L^\infty([-1,1])} = \sqrt{2j+1}$ for example leads to $m = s$. For some choices of ω the admissible indices form sets like hyperbolic crosses that are well-known in approximation theory [6, Section 1.7]. If the sequence is chosen appropriately, the dimension of $\mathcal{V}_m$ may even be independent from the spatial dimension d [18, Remark 5.5].

If the sequence is non-increasing then $\mathcal{V}_m$ as defined above may not be finite-dimensional. In this case however the covering number of $\mathcal{M}_{\omega,s}$ is infinite as well. Therefore, the intersection with an artificially chosen finite dimensional $\mathcal{V}_m$ is necessary.

Lemma 3.13. *Let $\mathcal{V}_m$ be an m-dimensional subspace spanned by a subset of $\{\mathbf{B}_j\}_{j\in\mathbb{N}}$ and consider the model set $\mathcal{M}_{\omega,s}\cap\mathcal{V}_m$. Then*

$$\hat{b}(x) \leq s\frac{\|\Omega^{-1}\mathbf{B}(x)\|_2^4}{\|\Omega^{-1}\mathbf{B}(x)\|_1^2} \quad \textit{with } \Omega := \operatorname{diag}(\omega).$$

Proof. Observe that by the Cauchy-Schwarz inequality

$$\|\Omega\mathbf{v}\|_1 = \sum_{j\in\operatorname{supp}(\mathbf{v})} |\mathbf{v}_j|\omega_j \leq \|\mathbf{v}\|_{\omega,0}\|\mathbf{v}\|_2.$$

Defining the model set

$$\mathcal{M} := \{v\in\mathcal{V}_m : \|\Omega\mathbf{v}\|_1 \leq \sqrt{s}\|\mathbf{v}\|_2\}$$

we have the inclusion $\mathcal{M}_{\omega,s}\cap\mathcal{V}_m\subseteq\mathcal{M}$. Since we know that $\hat{b}$ for $\mathcal{M}_{\omega,s}\cap\mathcal{V}_m$ is bounded by $\hat{b}$ for $\mathcal{M}$, we derive an estimate for the larger set.

Recall that

$$\hat{b}(x) := \sup_{v\in\mathcal{M}}\frac{\mathbf{v}^\intercal G(x)\mathbf{v}}{\|\mathbf{v}\|_2^2} \quad \text{with } G(x) := \mathbf{B}(x)\mathbf{B}(x)^\intercal.$$

Since $\|\mathbf{v}\|_2^{-1}\leq\sqrt{s}\|\Omega\mathbf{v}\|_1^{-1}$ for all $v\in\mathcal{M}$, we derive the bound

$$\hat{b}(x) \leq s\sup_{v\in\mathcal{M}}\frac{\mathbf{v}^\intercal G(x)\mathbf{v}}{\|\Omega\mathbf{v}\|_1^2} \leq s\sup_{\substack{\mathbf{v}\in\mathbb{R}^m\\ \mathbf{w}=\Omega\mathbf{v}}}\frac{\mathbf{w}^\intercal\Omega^{-1}G(x)\Omega^{-1}\mathbf{w}}{\|\mathbf{w}\|_1^2} = s\frac{\|\Omega^{-1}\mathbf{B}(x)\|_2^4}{\|\Omega^{-1}\mathbf{B}(x)\|_1^2}.$$

■

According to Theorem 3.1 the sampling density and weight function can be chosen optimally for every given model set. For $\mathcal{M}_{\omega,s}$ however, this role is reversed. The model set $\mathcal{M}_{\omega,s}$ depends implicitly on the weight function and changing w may change the set. When w is replaced by w^{new} the weight sequence ω might need to be adapted to ensure $\omega_j^{\mathrm{new}} \geq \|\mathbf{B}_j\|_{w^{\mathrm{new}},\infty}$. This restricts the model class on indices where $\omega_j \leq \|\mathbf{B}_j\|_{w^{\mathrm{new}},\infty}$ but it may allow us to broaden the class for indices where $\|\mathbf{B}_j\|_{w_{\mathrm{new}},\infty} \leq \|\mathbf{B}_j\|_{w,\infty}$. This is illustrated in Figure 5. However, the simplest way to ensure this is by scaling ω uniformly by a constant. Independent from the dependence on the model class this reweighting results in a superior sampling density which is illustrated in Figure 6.

Remark 3.14. *Theorem 3.12 states a sample complexity of*

$$n \gtrsim s^2(s\ln(m) - s\ln(\delta) - \ln(1-p))\delta^{-2},$$

which depends only logarithmically on the ambient dimension. The result can be compared with Theorem 5.2 in [6] where

$$n \gtrsim s\max\{\ln^3(s)\ln(m), \ln(p^{-1})\}\delta^{-2}$$

or Theorems 4.4 and 8.4 in [19] where

$$n \gtrsim s\max\{\ln^2(s)\ln(m)\ln(n), \ln(p^{-1})\}\delta^{-2}\max_{j\in[m]}\|\mathbf{B}_j\|_{w,\infty}^2.$$

This shows that our bound is qualitatively similar to specialized bounds for this model class even though the developed approach is more general.

The theory presented in this subsection can be generalized easily to dictionary learning. This is stated without proof in the following theorem.

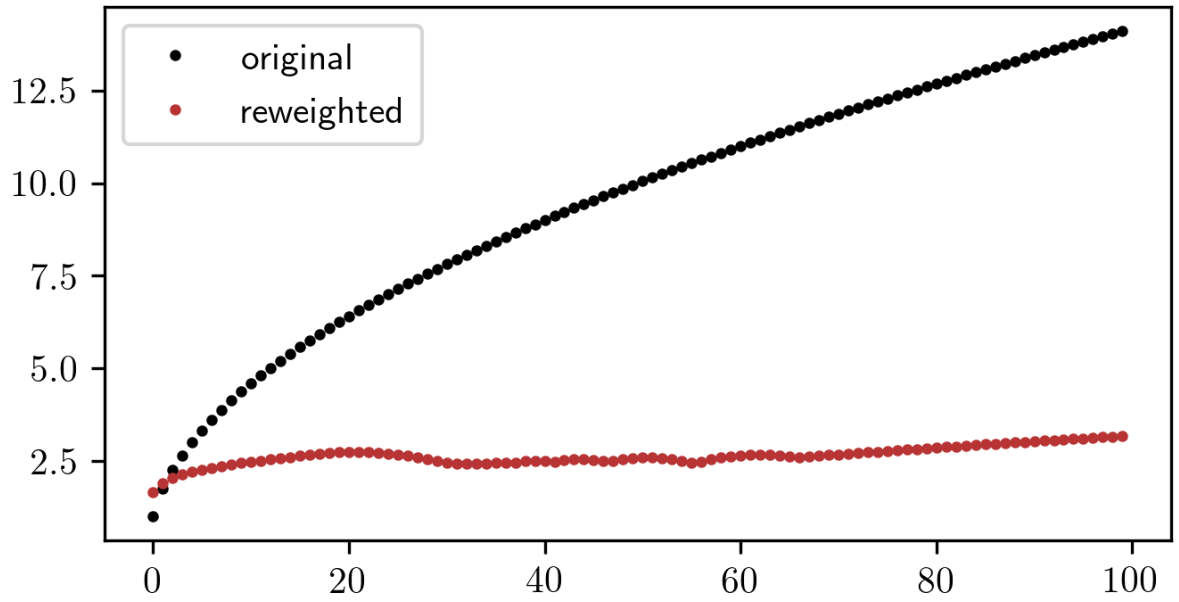


Figure 5. The weight sequences for the original weight function are bounded by $\omega_j \geq \|\mathbf{B}_j\|_{L^\infty}$ (black). The weight sequences for the new weight function are bounded by $\omega_j^{\mathrm{new}} \geq \|\mathbf{B}_j\|_{w^{\mathrm{new}},\infty}$ (red).

Theorem 3.15. *Assume that $\{\mathbf{B}_j\}_{j\in\mathbb{N}}$ is a Riesz sequence satisfying*

$$c\|\mathbf{v}\|_2^2 \leq \left\|\sum_{j\in\mathbb{N}} \mathbf{v}_j \mathbf{B}_j\right\|^2 \leq C\|\mathbf{v}\|_2^2$$

and redefine the model class

$$\mathcal{M}_{\omega,s} := \{v \in \mathcal{V} : \exists \mathbf{v} : v = \sum_{j=1}^{\infty} \mathbf{v}_j \mathbf{B}_j \wedge \|\mathbf{v}\|_{\omega,0} \leq s\}.$$

Let moreover $\mathcal{V}_m$ be an m-dimensional subspace spanned by a subset of $\{\mathbf{B}_j\}_{j\in\mathbb{N}}$. Then it holds that

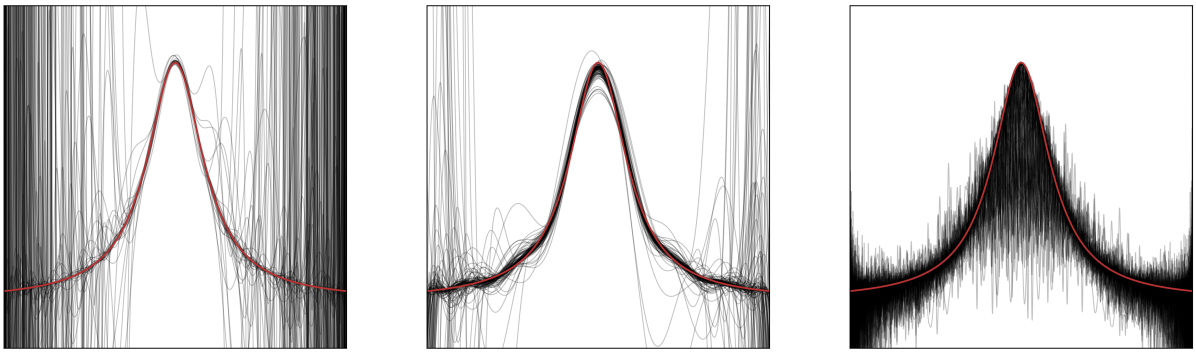
- $\mathcal{M}_{\omega,s} \subseteq \mathcal{M}_{\omega,t}$ *for* $s \leq t$,
- $\mathcal{M}_{\omega,s} = -\mathcal{M}_{\omega,s}$,
- $\mathcal{M}_{\omega,s} + \mathcal{M}_{\omega,t} \subseteq \mathcal{M}_{\omega,s+t}$,
- $\|v\|_{w,\infty} \leq \frac{\sqrt{s}}{c}\|v\|$ *for all* $v \in \mathcal{M}_{\omega,s}$ *and*
- $\nu_{\|\bullet\|_{w,\infty}}(U(\mathcal{M}_{\omega,s} \cap \mathcal{V}_m), r) \lesssim \left(\frac{Cm}{r\sqrt{s}}\right)^s$.

■

3.4. **Tensors of rank r.** For $M \in \mathbb{N}$ consider $\mathcal{V} := (L^2(Y,\rho))^{\otimes M}$ and define the m-dimensional subspace $\mathcal{V}_m \subseteq L^2(Y,\rho)$ spanned by the *orthonormal* basis functions $\{\mathbf{B}_j\}_{j\in[m]}$. The tensor product space $\mathcal{V}_m^{\otimes M} \subseteq \mathcal{V}$ is isomorphic to the space of coefficient tensors

$$\mathcal{V}_m^{\otimes M} \simeq (\mathbb{R}^m)^{\otimes M}.$$

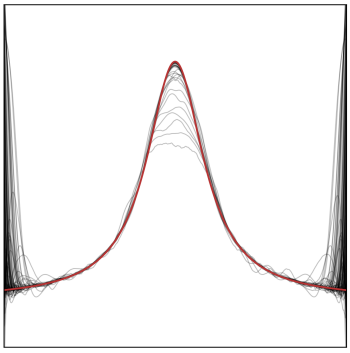
Due to the exponential scaling of the dimension m^M with respect to M, this space is infeasible for numerical computations. It has proven useful in applications to restrict the model class to functions



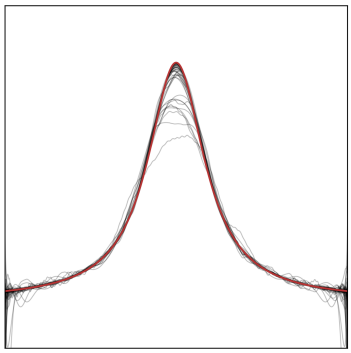
(a) Exact inversion

(b) Least squares

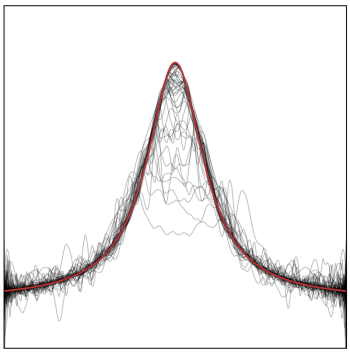
(c) ℓ^1, uniform ω



(d) ℓ^1, original ω



(e) Reweighted ℓ^1, original ω



(f) Reweighted ℓ^1, minimal ω

Figure 6. Overlaid interpolations of the function $f(x) = \frac{1}{1+25x^2}$ (red) by Legendre polynomials using various reconstruction methods. Different interpolations correspond to different random draws of $n = 30$ sampling points. The subfigures 6a and 6b show least squares estimates of the coefficients of the first 30 and 15 basis functions respectively. 6c displays the results of unweighted ℓ^1-minimization and 6d displays the results of weighted ℓ^1-minimization using the weight sequence $\omega_j = \|\mathbf{B}_j\|_{L^\infty}$. In all aforementioned cases the sampling points are drawn according to the uniform measure on $[-1, 1]$. The subplots 6f and 6e use samples that are drawn according to the optimal sampling density as given in Lemma 3.13. 6e uses the original sequence $\omega_j = \|\mathbf{B}_j\|_{L^\infty}$ while 6f uses the minimal possible weight sequence $\omega_j = \|\mathbf{B}_j\|_{w,\infty} \lesssim \|\mathbf{B}_j\|_{L^\infty}$.

$v \in \mathcal{V}_m^{\otimes M}$ for which the corresponding coefficient tensors $\mathbf{v} \in (\mathbb{R}^m)^{\otimes M}$ exhibit a small rank. Define the set of *tensors of CP-rank* r [20] recursively as

$$\begin{aligned} \mathcal{T}_1 &:= \{v \in \mathcal{V} : \mathbf{v} = \mathbf{v}_1 \otimes \cdots \otimes \mathbf{v}_M \text{ with } \mathbf{v}_1, \dots, \mathbf{v}_M \in \mathbb{R}^m\}, \\ \mathcal{T}_{r+1} &:= \mathcal{T}_r + \mathcal{T}_1. \end{aligned}$$

We compute the variation constant for this decomposition in the subsequent theorem. Note that the statement of this theorem is not constrained to the constant weight functions $w \equiv 1$ and the class of canonical tensor decompositions $\mathcal{T}_r$. In fact, it holds for any weight function $w \in \mathcal{T}_1$ and for any class of tensor decompositions $\mathcal{M}$ that satisfies $\mathcal{T}_1 \subseteq \mathcal{M}$. This includes all tree-shaped tensor formats including the Tucker format, the tensor-train (TT) format and general hierarchical tensor formats [21–23].

Theorem 3.16. *Assume that* $w \equiv 1$. *Then*

$$K(U(\mathcal{T}_r)) = K(U(\mathcal{V}_m))^M = K(U(\mathcal{V}_m^{\otimes M})).$$

Proof. We first prove the second equality. For this recall that $\{\mathbf{B}_j\}_{j\in[m]}$ is an *orthonormal* basis for $\mathcal{V}_m$ and define the *orthonormal* tensor product basis

$$\mathbf{B}_\alpha^{\otimes M}(y) = \prod_{j=1}^M \mathbf{B}_{\alpha_j}(y_j).$$

Using Example 3.3 we derive

$$\begin{aligned}
K(U(\mathcal{V}_m^{\otimes M})) = K_{\mathbf{B}^{\otimes M},w} &= \operatorname*{ess\,sup}_{y\in Y^M} \sum_{\alpha\in[m]^M} \mathbf{B}_\alpha^{\otimes M}(y)^2 \\
&= \operatorname*{ess\,sup}_{y\in Y^M} \sum_{\alpha\in[m]^M} \prod_{j=1}^M \mathbf{B}_{\alpha_j}(y_j)^2 \\
&= \operatorname*{ess\,sup}_{y\in Y^M} \prod_{j=1}^M \sum_{\alpha_j\in[m]} \mathbf{B}_{\alpha_j}(y_j)^2 \\
&= \prod_{j=1}^M \operatorname*{ess\,sup}_{y_j\in Y} \sum_{\alpha_j\in[m]} \mathbf{B}_{\alpha_j}(y_j)^2 \\
&= K(U(\mathcal{V}_m))^M.
\end{aligned}$$

To prove the first equality, observe that $\mathcal{T}_r \subseteq \mathcal{T}_{r+1}$ and $\mathcal{T}_{m^M} = \mathcal{V}_m^{\otimes M}$ and consequently $K(U(\mathcal{T}_1)) \leq K(U(\mathcal{T}_r)) \leq K(U(\mathcal{V}_m^{\otimes M}))$. The assertion now follows directly from the definition of $K(U(\mathcal{T}_1))$ by

$$K(U(\mathcal{T}_1)) = \sup_{\substack{u_j\in\mathcal{V}_m \\ \|u_j\|=1}} \operatorname*{ess\,sup}_{y_j\in Y} \prod_{j=1}^d |u_j(y_j)|^2 = \prod_{j=1}^d K(U(\mathcal{V}_m)) = K(U(\mathcal{V}_m))^M.$$

■

We cannot expect better results for such a general class of functions. Without additional regularity assumptions, every factor of a rank-1 tensor product can become arbitrarily large. This can be condensed into the statement that regularity may induce low-rank structure but low-rank structure does not provide regularity, cf. [24]. We are not aware of better estimates of the variation constant in this setting and would like to point out the L_∞ estimates in [17].

Despite this unfavourable result, the present theory notably can be used in this setting. Corollary 2.9 can be applied by utilizing the bound $\|\bullet\|_{w,\infty} \leq \sqrt{K}\|\bullet\|$ together with the isometry $\|\bullet\| = \|\bullet\|_2$ and bounds for the covering number of low-rank tensor formats, see e.g. [7]. To the knowledge of the authors this is the first estimate for the sample complexity of this function class in the setting of continuous approximation.

4. Connection with Empirical Risk Minimisation

In *empirical risk minimisation* (ERM) as considered in [1], an empirical risk functional

$$\mathcal{J}_n(v) := \frac{1}{n}\sum_{i=1}^n \ell(v, y_i) \approx \mathbb{E}_Y[\ell(v, Y)] =: \mathcal{J}(v)$$

is minimised. Define the minimizers

$$v^* := \arg\min_{v\in\mathcal{V}} \mathcal{J}(v),\ v^*_{\mathcal{M}} := \arg\min_{v\in\mathcal{M}} \mathcal{J}(v),\ v^*_{\mathcal{M},n} := \arg\min_{v\in\mathcal{M}} \mathcal{J}_n(v).$$

If $\mathcal{M}$ is bounded, $\mathcal{J}$ is Lipschitz continuous on $\mathcal{M}$ as well as Lipschitz smooth[1] and strongly convex on $\mathcal{V}$, it is shown in [1] that the empirical optimum is quasi-optimal with high probability, i.e.[2]

$$\|v^* - v^*_{\mathcal{M},n}\| \lesssim \|(1-P)v^*\|.$$

Using $\mathrm{RIP}_{Pv^*-\mathcal{M}}(\delta)$ and $\mathrm{RIP}_{\{(1-P)v^*\}}(\delta)$, we can employ Theorem 2.12 to derive similar estimate under weaker assumptions.

First, assume that $\mathcal{J}_n$ is Lipschitz smooth and strongly convex on $\mathcal{V}$ and that $v^* \in \arg\min_{v\in\mathcal{V}} \mathcal{J}_n(v)$. Then, for all $v \in \mathcal{V}$

$$\|v - v^*\|_n^2 \overset{(4a)}{\lesssim} \mathcal{J}_n(v) - \mathcal{J}_n(v^*) \overset{(4b)}{\lesssim} \|v - v^*\|_n^2.$$

Here, (4a) comes from strong convexity and (4b) from Lispschitz smoothness. By the triangle inequality and $\mathrm{RIP}_{Pv^*-\mathcal{M}}$,

$$\begin{aligned}\|v^*_{\mathcal{M},n} - v^*\|^2 &\lesssim \|v^*_{\mathcal{M},n} - Pv^*\|^2 + \|(P-1)v^*\|^2\\ &\lesssim \|v^*_{\mathcal{M},n} - Pv^*\|_n^2 + \|(P-1)v^*\|^2\end{aligned}$$

Another triangle inequality and (4a) yield

$$\begin{aligned}\|v^*_{\mathcal{M},n} - Pv^*\|_n^2 &\lesssim \|v^*_{\mathcal{M},n} - v^*\|_n^2 + \|(1-P)v^*\|_n^2\\ &\lesssim \mathcal{J}_n(v^*_{\mathcal{M},n}) - \mathcal{J}_n(v^*) + \|(P-1)v^*\|_n^2.\end{aligned}$$

Recalling the definition of $v^*_{\mathcal{M},n}$ and using (4b) leads to

$$\begin{aligned}\mathcal{J}_n(v^*_{\mathcal{M},n}) - \mathcal{J}_n(v^*) &\leq \mathcal{J}_n(Pv^*) - \mathcal{J}_n(v^*)\\ &\lesssim \|(P-1)v^*\|_n^2.\end{aligned}$$

Using $\mathrm{RIP}_{\{(1-P)v^*\}}$ and combining the preceding estimates, we obtain

$$\|v^* - v^*_{\mathcal{M},n}\| \lesssim \|(1-P)v^*\|.$$

Second, assume that $\mathcal{J}$ is Lipschitz smooth and strongly convex on $\mathcal{V}$. Then, for all $v \in \mathcal{V}$

$$\|v - v^*\|^2 \overset{(4a)}{\lesssim} \mathcal{J}(v) - \mathcal{J}(v^*) \overset{(4b)}{\lesssim} \|v - v^*\|^2.$$

Using (4b), $\mathrm{RIP}_{Pv^*-\mathcal{M}}(\delta)$ and $\mathrm{RIP}_{\{(1-P)v^*\}}(\delta)$ with Theorem 2.12 yields

$$\begin{aligned}\mathcal{J}(P_n v^*) - \mathcal{J}(v^*) &\lesssim \|(P_n - 1)v^*\|^2\\ &\lesssim \|(P-1)v^*\|^2.\end{aligned}$$

Then, with the definition of $v^*_{\mathcal{M}}$ and (4a), we obtain

$$\begin{aligned}\mathcal{J}(P_n v^*) - \mathcal{J}(v^*) &\lesssim \|v^*_{\mathcal{M}} - v^*\|^2\\ &\lesssim \mathcal{J}(v^*_{\mathcal{M}}) - \mathcal{J}(v^*).\end{aligned}$$

[1] i.e. its gradient is Lipschitz continuous

[2] Note that we hide multiplicative constants in "$\lesssim$".

This means that we can approximate v^* in two ways: either by minimizing the norm $\|v^* - \bullet\|$ or by minimizing a more general cost functional $\mathcal{J}$. In both cases, we obtain a quasi-optimal bound.

Remark 4.1. *In the age of artificial neural networks, least-squares methods are used with nonlinear model classes or replaced by the more general empirical risk minimization (ERM). When the model class is bounded, the ERM approach allows for a similar convergence bound as least squares methods which for instance was recently analysed and demonstrated in [1]. The probability of this bound however decreases exponentially with the best approximation error $\|(1-P)u\|$ and vanishes when the best approximation error is zero.*

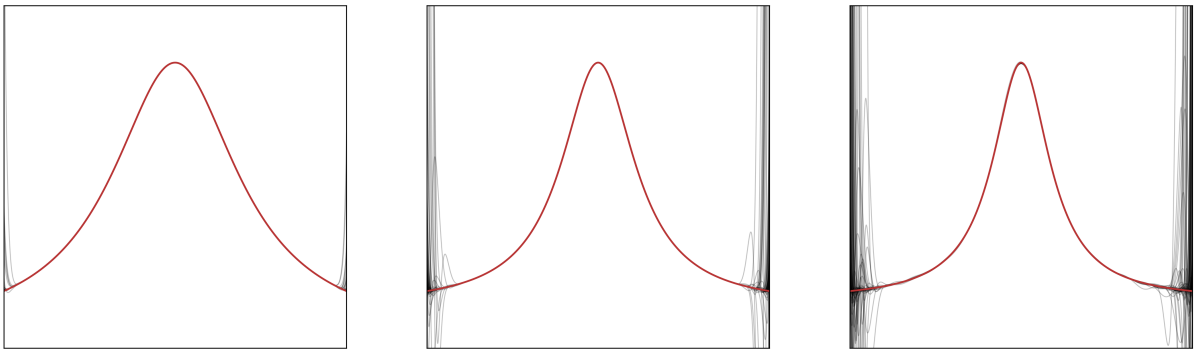
5. Discussion

In Section 2 we derive error bounds for the empirical approximation (1) by utilizing a *restricted isometry property (RIP)* that is defined for general (nonlinear) model classes. When the number of samples is sufficiently large, this RIP holds with high probability and the resulting approximation is quasi-optimal. The required number of samples depends mainly on the variation constant K of the model class and at most linearly on the ambient dimension. With respect to the number of samples we observe exponential convergence of the expected error until the RIP is satisfied almost surely. After that the convergence transitions to a (sub-)linear rate.

In Sections 3.1 and 3.3 we apply our central theory developed in Section 2 to the model classes of linear function spaces and sets of functions with sparse representation. For both cases we derive results that are qualitatively similar to known specialized results for the respective model classes. We assume that these bounds can be tightened by using more advanced techniques (cf. [25, 26]) but expect the bounds to improve by a polynomial factor at most. Moreover, we investigate improved convergence rates of the empirical approximation when stronger norms are used. By bounding the variation constant on subsets of Sobolev spaces, we provide a theoretical reasoning for this effect.

In Section 3.4 we derive the first bound on the sample complexity of low-rank tensors in the setting of nonlinear least squares. We however also observe that the variation constant of this model set is equal to that of the ambient space and that additional regularity assumptions are advisable when optimizing in this class. As an example, we refer to [27] where the authors allow only local interactions of the component tensors in the model class of TT tensors.

The fact that this model class has such an unfavorable variation constant is especially surprising since one goal of this work has been to reconcile observations from the very promising experiments in [1] with the theory presented therein. In [1] we reconstructed a function u mapping from $\mathbb{R}^M$ to a Hilbert space $\mathcal{X}$ and observed that the numerical convergence significantly exceeded the theoretical predictions. A low-rank tensor ansatz similar to the one described in Section 3.4 was employed. The basis was constituted of tensorized Hermite polynomials and the samples were drawn according to the standard Gaussian measure. Under these conditions, the present theory predicts the RIP to fail almost certainly. Nevertheless, the observed convergence rate was much higher than the Monte Carlo rate which is most certainly due to the regularity of the approximated function u. We expect that this is because Theorem 2.12 requires the RIP to hold for the shifted model class $u - \mathcal{M}$ where the additional regularity of u may reduce the variation constant. Figure 7 illustrates this behaviour. The model class used for all three experiments is the same and only the regularity of the function varies. Event though the best approximation error in all three cases is bounded by 10^{-3} we can observe how the empirical



(a) $f(x) = \frac{1}{1+5x^2}$ **(b)** $f(x) = \frac{1}{1+15x^2}$ **(c)** $f(x) = \frac{1}{1+25x^2}$

Figure 7. Overlaid least squares approximations of the function $f(x) = \frac{1}{1+cx^2}$ (red) by Legendre polynomials of degree 29. Different approximations correspond to different random draws of $n = 100$ sampling points from the uniform measure on $[-1, 1]$.

approximations deteriorate with decreasing regularity. The relative errors for the empirical approximation increase from 10^{-2} to 10^{1}. This phenomenon will be investigated in future research.

Finally, in Section 4 we briefly describe the connection of this work to *empirical risk minimization (ERM)* as scrutinized in [1]. We show that under certain convexity and smoothness assumptions the fast convergence rates derived in Section 2 carry over to the ERM setting. This, for example, enables to apply our theory to solve high-dimensional elliptic PDEs by minimizing the respective Dirichlet energy.

References

[1] M. Eigel, R. Schneider, P. Trunschke, and S. Wolf. “Variational Monte Carlo—bridging concepts of machine learning and high-dimensional partial differential equations”. In: *Advances in Computational Mathematics* (10/2019).

[2] V. N. Vapnik and A. Y. Chervonenkis. “Necessary and Sufficient Conditions for the Uniform Convergence of Means to their Expectations”. In: *Theory of Probability & Its Applications* 26.3 (01/1982), pp. 532–553.

[3] F. Cucker and D. X. Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2007.

[4] E. J. Candès, J. K. Romberg, and T. Tao. “Stable signal recovery from incomplete and inaccurate measurements”. In: *Communications on Pure and Applied Mathematics* 59.8 (2006), pp. 1207–1223. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpa.20124.

[5] Y. C. Eldar and G. Kutyniok. *Compressed sensing: theory and applications*. Cambridge university press, 2012.

[6] H. Rauhut and R. Ward. “Interpolation via weighted $\ell 1$ minimization”. In: *Applied and Computational Harmonic Analysis* 40.2 (03/2016), pp. 321–351.

[7] H. Rauhut, R. Schneider, and Ž. Stojanac. “Low rank tensor recovery via iterative hard thresholding”. In: *Linear Algebra and its Applications* 523 (2017), pp. 220–262.

[8] L. Grasedyck and S. Krämer.
"Stable ALS approximation in the TT-format for rank-adaptive tensor completion".
In: *Numerische Mathematik* 143.4 (08/2019), pp. 855–904.

[9] J. Berner, P. Grohs, and A. Jentzen. *Analysis of the generalization error: Empirical risk minimization over deep artificial neural networks overcomes the curse of dimensionality in the numerical approximation of Black-Scholes partial differential equations*. 09/2018.

[10] G. Kutyniok, P. Petersen, M. Raslan, and R. Schneider.
A Theoretical Analysis of Deep Neural Networks and Parametric PDEs. 04/2019.

[11] A. Cohen and G. Migliorati. *Optimal weighted least-squares methods*. 2016.
eprint: arXiv:1608.00512.

[12] B. Bohn. "On the convergence rate of sparse grid least squares regression".
In: *Sparse Grids and Applications-Miami 2016*. Springer, 2018, pp. 19–41.

[13] Y. Traonmilin and R. Gribonval.
"Stable recovery of low-dimensional cones in Hilbert spaces: One RIP to rule them all".
In: *Applied and Computational Harmonic Analysis* 45.1 (07/2018), pp. 170–205.

[14] R. Vershynin. "On the role of sparsity in Compressed Sensing and random matrix theory".
In: *2009 3rd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)* (12/2009).

[15] H. Gerlach and H. von der Mosel. "On Sphere-Filling Ropes".
In: *The American Mathematical Monthly* 118 (05/2010).

[16] E. Kowalski.
"POINTWISE BOUNDS FOR ORTHONORMAL BASIS ELEMENTS IN HILBERT SPACES". In: 2011.

[17] W. Hackbusch. "L_∞-estimation of tensor truncations".
In: *Numer. Math.* 125 (2013), pp. 419–440.

[18] H. Rauhut and C. Schwab. "Compressive sensing Petrov-Galerkin approximation of high-dimensional parametric operator equations".
In: *Mathematics of Computation* 86.304 (05/2016), pp. 661–700.

[19] H. Rauhut. "Compressive sensing and structured random matrices".
In: *Theor Found Numer Methods Sparse Recover* 9 (01/2010).

[20] F. L. Hitchcock. "The Expression of a Tensor or a Polyadic as a Sum of Products".
In: *Journal of Mathematics and Physics* 6.1-4 (1927), pp. 164–189. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sapm192761164.

[21] M. Bachmayr and R. Schneider.
"Iterative methods based on soft thresholding of hierarchical tensors".
In: *Foundations of Computational Mathematics* 17.4 (2017), pp. 1037–1083.

[22] W. Hackbusch. *Tensor spaces and numerical tensor calculus*. Vol. 42.
Springer Science & Business Media, 2012.

[23] L. Grasedyck and W. Hackbusch.
"An introduction to hierarchical (H-) rank and TT-rank of tensors with examples".
In: *Computational Methods in Applied Mathematics Comput. Methods Appl. Math.* 11.3 (2011), pp. 291–304.

[24] M. Griebel and H. Harbrecht. *Analysis of tensor approximation schemes for continuous functions*. 2019. arXiv: 1903.04234 [math.NA].

[25] R. Vershynin. *High-Dimensional Probability*. Cambridge University Press, 09/2018.

[26] S. Dirksen. “Tail bounds via generic chaining”. In: *Electronic Journal of Probability* 20.0 (2015).
[27] A. Goeßmann, M. Götte, I. Roth, R. Sweke, G. Kutyniok, and J. Eisert. *Tensor network approaches for learning non-linear dynamical laws*. 2020. arXiv: 2002.12388 [math.NA].
[28] V. Burenkov. “Extension theorems for Sobolev spaces”. In: *The Maz’ya Anniversary Collection*. Ed. by J. Rossmann, P. Takáč, and G. Wildenhain. Basel: Birkhäuser Basel, 1999, pp. 187–200.
[29] E. Novak, M. Ullrich, H. Woźniakowski, and S. Zhang. “Reproducing kernels of Sobolev spaces on $\mathbb{R}^d$ and applications to embedding constants and tractability”. In: *Analysis and Applications* 16.05 (2018), pp. 693–715.
[30] I. S. Gradshteyn, I. M. Ryzhik, and D. F. Hays. “Table of Integrals, Series, and Products”. In: 1943.

Appendix A. Proof of Lemma 2.7

The proof consists of two steps. In the first step we derive Lemma A.3 to show that there exists $\nu \in \mathbb{N}$ and $\{u_j\}_{j\in[\nu]} \subseteq U(A)$ such that

$$\mathbb{P}\Big[\sup_{u\in U(A)} |\|u\|^2 - \|u\|_n^2| > \delta\Big] \leq \mathbb{P}\Big[\max_{1\leq j\leq \nu} |\|u_j\|^2 - \|u_j\|_n^2| > \tfrac{\delta}{2}\Big].$$

Using a union bound argument it follows that

$$\begin{aligned}\mathbb{P}\Big[\max_{1\leq j\leq \nu} |\|u_j\|^2 - \|u_j\|_n^2| > \tfrac{\delta}{2}\Big] &\leq \sum_{1\leq j\leq \nu} \mathbb{P}\Big[|\|u_j\|^2 - \|u_j\|_n^2| > \tfrac{\delta}{2}\Big] \\ &\leq \nu \max_{1\leq j\leq \nu} \mathbb{P}\Big[|\|u_j\|^2 - \|u_j\|_n^2| > \tfrac{\delta}{2}\Big].\end{aligned}$$

In the second step we prove Lemma A.5 which allows us to bound the probability

$$\mathbb{P}\Big[|\|u_j\|^2 - \|u_j\|_n^2| > \tfrac{\delta}{2}\Big] \leq 2\exp(-\tfrac{\delta^2 n}{2K^2})$$

for each $1 \leq j \leq \nu$ by a standard concentration inequality. Combining both inequalities yields the statement.

In the following we are concerned with proving Lemmas A.3 and A.5 which both rely on properties of the function $\ell_y : u \mapsto w(y)|u|_y^2$.

Lemma A.1. *The function $\ell_y : u \mapsto w(y)|u|_y^2$ has the properties*

- $|\ell_y(u)| \leq K$ *and*
- $|\ell_y(u) - \ell_y(v)| \leq 2\sqrt{K}\|u - v\|_{w,\infty}$

for all $u, v \in U(A)$.

Proof. Let $u, v \in U(A)$. The first statement follows immediately by

$$|\ell_y(u)| \leq \sup_{u\in U(A)} \operatorname*{ess\,sup}_{y\in Y} w(y)|u|_y^2 = K.$$

To prove the second statement we consider the seminorm $\hbar_y := \sqrt{\ell_y}$ and use the reverse triangle inequality

$$|\hbar_y(u) - \hbar_y(v)| \leq \hbar_y(u - v) \leq \operatorname*{ess\,sup}_{y\in Y} \hbar_y(u - v) = \|u - v\|_{w,\infty}.$$

Since $\hbar_y$ is bounded by $\sqrt{K}$, we can use the Lipschitz continuity of $x \mapsto x^2$ on $[-\sqrt{K}, \sqrt{K}]$ to conclude

$$|\ell_y(u) - \ell_y(v)| \leq 2\sqrt{K}|\hbar_y(u) - \hbar_y(u)| \leq 2\sqrt{K}\|u - v\|_{w,\infty}. \qquad \square$$

As an intermediate step we first prove Lemma A.2 from which Lemma A.3 follows almost immediately.

Lemma A.2. *Let $\nu := \nu_{\|\bullet\|_{w,\infty}}\big(U(A), \frac{\delta}{8\sqrt{K}}\big)$ and $\{u_j\}_{j\in[\nu]}$ be the centres of the corresponding covering. Then almost surely*

$$\sup_{u\in U(A)} |\|u\|^2 - \|u\|_n^2| \leq \tfrac{\delta}{2} + \max_{1\leq j\leq \nu} \big|\|u_j\|^2 - \|u_j\|_n^2\big|.$$

Proof. Let $u \in U(A)$ be given. Then by definition of the $\{u_j\}_{j\in[\nu]}$, there is a specific u_j with $\|u - u_j\|_{w,\infty} \leq \frac{\delta}{8\sqrt{K}}$. By Lemma A.1 and Jensen's inequality we know that

$$|\|u\|^2 - \|u_j\|^2| \leq \int_Y |\ell_y(u) - \ell_y(u_j)| \,\mathrm{d}\rho(y) \leq 2\sqrt{K}\|u - v\|_{w,\infty} \leq \tfrac{\delta}{4}$$

and almost surely

$$|\|u\|_n^2 - \|u_j\|_n^2| \leq \tfrac{1}{n}\sum_{i=1}^n |\ell_{y_i}(u) - \ell_{y_i}(u_j)| \leq 2\sqrt{K}\|u - v\|_{w,\infty} \leq \tfrac{\delta}{4}.$$

Therefore, by triangle inequality,

$$\begin{aligned} |\|u\|^2 - \|u\|_n^2| &\leq |\|u\|^2 - \|u\|_n^2 - (\|u_j\|^2 - \|u_j\|_n^2)| + |\|u_j\|^2 - \|u_j\|_n^2| \\ &\leq |\|u\|^2 - \|u_j\|^2| + |\|u\|_n^2 - \|u_j\|_n^2| + |\|u_j\|^2 - \|u_j\|_n^2| \\ &\leq \tfrac{\delta}{2} + |\|u_j\|^2 - \|u_j\|_n^2| \quad \text{almost surely.} \end{aligned}$$

Taking the maximum concludes the proof. $\square$

Lemma A.3. *Let $\nu := \nu_{\|\bullet\|_{w,\infty}}\big(U(A), \frac{\delta}{8\sqrt{K}}\big)$ and $\{u_j\}_{j\in[\nu]}$ be the centres of the corresponding covering. Then*

$$\mathbb{P}\Big[\sup_{u\in U(A)} |\|u\|^2 - \|u\|_n^2| > \delta\Big] \leq \mathbb{P}\Big[\max_{1\leq j\leq\nu} |\|u_j\|^2 - \|u_j\|_n^2| > \tfrac{\delta}{2}\Big].$$

Proof. By Lemma A.2

$$\sup_{u\in U(A)} |\|u\|^2 - \|u\|_n^2| \leq \tfrac{\delta}{2} + \max_{1\leq j\leq \nu} \big|\|u_j\|^2 - \|u_j\|_n^2\big|$$

holds almost surely. In this event we know that

$$\sup_{u\in U(A)} |\|u\|^2 - \|u\|_n^2| > \delta \Rightarrow \max_{1\leq j\leq\nu} |\|u_j\|^2 - \|u_j\|_n^2| > \tfrac{\delta}{2}$$

which concludes the proof. $\square$

To prove Lemma A.5 we first recall a standard concentration result from statistics.

Lemma A.4 (Hoeffding 1963)**.** *Let $\{X_i\}_{i\in[N]}$ be a sequence of i.i.d. bounded random variables $|X_i| \leq M$ and define $\overline{X} := \frac{1}{N}\sum_{i=1}^N X_i$. Then*

$$\mathbb{P}\big[|\mathbb{E}[\overline{X}] - \overline{X}| \geq \delta\big] \leq 2\exp\big(-\tfrac{2\delta^2 N}{M^2}\big).$$

The proof of Lemma A.5 is now a mere application of this result.

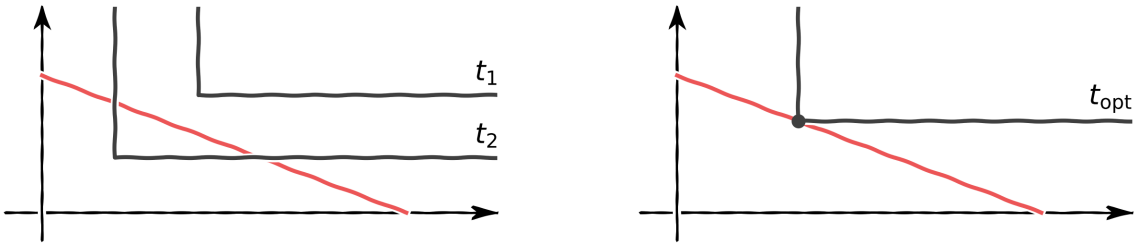


Figure 8. The set of feasible vectors v satisfying $v > 0$ and $\int_X \hat{b} v \,\mathrm{d}\rho = 1$ is displayed in red. The contour lines $\|\frac{1}{v}\|_{L^\infty(Y)} = t_{1/2/\text{opt}}$ for $t_1 < t_2$ (left) and for the optimal value t_{opt} (right) are drawn in black.

Lemma A.5. *Let $u_j \in U(A)$ then*

$$\mathbb{P}\Big[\big|\|u_j\|^2 - \|u_j\|_n^2\big| > \tfrac{\delta}{2}\Big] \leq 2\exp(-\tfrac{n\delta^2}{2K^2}).$$

Proof of Lemma A.5. The statement follows from an application of Lemma A.4 to the sequence of random variables $\{\ell_{y_i}(u_j)\}_{i=1}^n$. Since the samples y_i are i.i.d. the random variables $\ell_{y_i}(u)$ are i.i.d. as well. Moreover, by Lemma A.1 the variables are bounded in absolute value by K. Therefore, the assumptions for Lemma A.4 are satisfied. □

Appendix B. Proof of Theorem 3.1

To prove the first assertion it suffices to show that $\hat{b}$ is measurable. For this let $\{u_j\}_{j=1}^\infty$ be a countable dense subset in $\mathcal{M}$. Then

$$\hat{b}(y) := \sup_{u\in\mathcal{M}} |u|_y^2 = \sup_{j\in\mathbb{N}} |u_j|_y^2$$

is the supremum over a countable set of measurable functions and as such it is measurable.

We only sketch the proof of the second assertion. By substituting $w = (v\hat{b})^{-1}$, the minimization problem

$$\min_w K_w \quad \text{s.t.} \quad w \geq 0 \text{ and } \|w^{-1}\|_{L^1(Y,\rho)} = 1$$

is equivalent to

$$\min_v \|v^{-1}\|_{L^\infty(Y)} \quad \text{s.t.} \quad v > 0 \text{ and } \int_Y \hat{b} v \,\mathrm{d}\rho = 1,$$

which is a non-convex optimization problem under linear constraints. The assertion is then equivalent to the statement that the minimal v is a constant function. Figure 8 illustrates why this must be the case. The function v is split as $v = \alpha_1 v_1 + \alpha_2 v_2$ with $v_1, v_2 > 0$ having disjoint support and $\|v_1^{-1}\|_{L^\infty(Y)} = \|v_2^{-1}\|_{L^\infty(Y)} = 1$. Then $\|v^{-1}\|_{L^\infty(Y)} = \|(\alpha_1 v_1)^{-1} + (\alpha_2 v_2)^{-1}\|_{L^\infty(Y)} = \alpha_1^{-1} \vee \alpha_2^{-1}$.

Appendix C. Proof of Example 3.4

Recall that $\mathcal{V} := H^m(Y,\rho)$ where $Y \subseteq \mathbb{R}^d$ is a Lipschitz domain and $A \subseteq \mathcal{H} := H^M(Y,\rho)$ with $\ell := M - m > \frac{d}{2}$. It was shown in [28] that since Y is Lipschitz $H^m(Y)$ can be embedded isometrically into $H^m(\mathbb{R}^d)$. This means that we can restrict our analysis to the case $Y = \mathbb{R}^d$. Since

$\ell > \frac{d}{2}$, the Sobolev embedding theorem ensures that $D^\alpha v \in H^\ell(Y,\rho) \subseteq C^0(Y)$. This means that the seminorm of $H^m(Y,\rho)$ can be represented by

$$|v|_y^2 = \sum_{|\alpha|\le m} |[D^\alpha v](y)|^2 = \sum_{|\alpha|\le m} |L_y^\alpha v|^2$$

with the family of linear operators $L_y^\alpha : v \mapsto [D^\alpha v](y)$. In the following we compute

$$\kappa(y) = \|L_y\|^2_{\mathcal{L}(\mathcal{H},\mathbb{R}^{|\{|\alpha|\le m\}|})} = \sum_{|\alpha|\le m} \|L_y^\alpha\|\mathcal{H}^{*2}.$$

As in [29] the Riesz representative of L_y^α

$$K_y^\alpha(x) := \int_{\mathbb{R}^d} \frac{\prod_{j=1}^d (2\pi \mathrm{i} u_j)^{\alpha_j} \exp(2\pi \mathrm{i}(x-y)\cdot u)}{\sum_{|\beta|\le m+l} \prod_{j=1}^d (2\pi u_j)^{2\beta_j}} \,\mathrm{d}u$$

can be obtained by using the Fourier transform and some standard properties. Thus,

$$\begin{aligned}\|L_y^\alpha\|^2_{\mathcal{H}^*} = \|K_y^\alpha\|^2_{\mathcal{H}} = \langle K_y^\alpha, \overline{K_y^\alpha}\rangle_{\mathcal{H}} = [D^\alpha \overline{K_y^\alpha}](y)\\ = \int_{\mathbb{R}^d} \frac{\prod_{j=1}^d (2\pi u_j)^{2\alpha_j}}{\sum_{|\beta|\le m+l} \prod_{j=1}^d (2\pi u_j)^{2\beta_j}} \,\mathrm{d}u.\end{aligned}$$

By the change of variables $t_j = 2\pi u_j$

$$\|L_y^\alpha\|^2_{\mathcal{H}^*} = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{\prod_{j=1}^d t_j^{2\alpha_j}}{\sum_{|\beta|\le m+l} \prod_{j=1}^d t_j^{2\beta_j}} \,\mathrm{d}t.$$

The multinomial theorem states that

$$(1+\|t\|_2^2)^m = \sum_{|\alpha|\le m} \binom{m}{\alpha} \prod_{j=1}^d t_j^{2\alpha_j}.$$

As a consequence,

$$\sum_{|\alpha|\le m} \prod_{j=1}^d t_j^{2\alpha_j} \le (1+\|t\|_2^2)^m \le \Gamma(m+1) \sum_{|\alpha|\le m} \prod_{j=1}^d t_j^{2\alpha_j}.$$

This leads to the estimate

$$\begin{aligned}\sum_{|\alpha|\le m} \|L_y^\alpha\|^2_{\mathcal{H}^*} &= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{\sum_{|\alpha|\le m} \prod_{j=1}^d t_j^{2\alpha_j}}{\sum_{|\beta|\le m+\ell} \prod_{j=1}^d t_j^{2\beta_j}} \,\mathrm{d}t\\ &\le \frac{\Gamma(m+\ell+1)}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{(1+\|t\|_2^2)^m}{(1+\|t\|_2^2)^{m+\ell}} \,\mathrm{d}t\\ &= \frac{\Gamma(m+\ell+1)}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{\mathrm{d}t}{(1+\|t\|_2^2)^\ell}\\ &= \frac{\Gamma(m+\ell+1)}{(2\pi)^d} \frac{2\pi^{d/2}}{\Gamma(\frac{d}{2})} \int_0^\infty \frac{s^{d-1}}{(1+s^2)^\ell} \,\mathrm{d}s.\end{aligned}$$

The recurrence relation (2.147) in [30] together with $ell > \frac{d}{2}$ yields

$$\int_0^\infty \frac{s^{d-1}}{(1+s^2)^l} \,\mathrm{d}s = \frac{d-2}{2\ell-d} \int_0^\infty \frac{s^{d-3}}{(1+s^2)^\ell} \,\mathrm{d}s = \ldots = \frac{\Gamma(\ell-\frac{d}{2})\Gamma(\frac{d}{2})}{2\Gamma(\ell)}.$$

Consequently,

$$\kappa(y) \leq (2\sqrt{\pi})^{-d} \frac{\Gamma(m+\ell+1)\Gamma(\ell-\frac{d}{2})}{\Gamma(\ell)}.$$