ESAIM: M2AN 56 (2022) 79–104
https://doi.org/10.1051/m2an/2021070

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

CONVERGENCE BOUNDS FOR EMPIRICAL NONLINEAR LEAST-SQUARES

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

Abstract. We consider best approximation problems in a nonlinear subset $\mathscr{M}$ of a Banach space of functions $(\mathscr{V}, \|\bullet\|)$. 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 $v \in \mathscr{M}$ of an unknown function $u \in \mathscr{V}$ by minimizing the empirical norm $\|u - v\|_n$. We consider this problem 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 is independent of the specified nonlinear least squares setting. Several model classes are examined and the analytical statements about the RIP are compared to existing sample complexity bounds from the literature. We find that for well-studied model classes our general bound is weaker but exhibits many of the same properties as these specialized bounds. Notably, we demonstrate the advantage of an optimal sampling density (as known for linear spaces) for sets of functions with sparse representations.

Mathematics Subject Classification. 62J02, 41A25, 41A65, 41A30.

Received April 2, 2020. Accepted October 20, 2021.

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 $(\mathscr{V}, \|\bullet\|)$ that can be well approximated in a given nonlinear subset (or *model class*) $\mathscr{M} \subseteq \mathscr{V}$. 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 $v \in \mathscr{V}$. This norm is a generalization of the $L^2(Y, \rho)$- and $H_0^1(Y, \rho)$-norms which are induced by the seminorms $|v|_y^2 = |v(y)|^2$ and $|v|_y^2 = \|\nabla v(y)\|_2^2$, respectively.

Keywords and phrases. Weighted nonlinear least squares, error analysis, convergence rates, weighted sparsity, tensor networks.

[1] Weierstrass Institute, Mohrenstrasse 39, D-10117 Berlin, Germany.
[2] TU Berlin, Straße des 17 Juni 136, D-10623 Berlin, Germany.
*Corresponding author: ptrunschke@mail.tu-berlin.de

We characterize any best approximation $u_{\mathscr{M}}$ in $\mathscr{M}$ by

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

In general, this approximation is not computable. We propose to approximate $u_{\mathscr{M}}$ by an estimator $u_{\mathscr{M},n}$ 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 can be chosen as any almost surely positive function $w > 0$ that satisfies $\int_Y w^{-1}\,\mathrm{d}\rho = 1$. Any corresponding *empirical* best approximation $u_{\mathscr{M},n}$ in $\mathscr{M}$ is characterized by

$$u_{\mathscr{M},n} \in \operatorname*{arg\,min}_{v\in\mathscr{M}} \|u - v\|_n. \tag{1.1}$$

Given this definition we can choose w such that the theoretical convergence rate of $\|u - u_{\mathscr{M},n}\| \xrightarrow{n\to\infty} \|u - u_{\mathscr{M}}\|$ is maximized. 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* (*cf.* [7]).

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

$$\|u - u_{\mathscr{M}}\| = \min_{v\in\mathscr{M}} \|u - v\|$$

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 $\|u - u_{\mathscr{M},n}\|$ is equivalent to this error with high probability.

Main result. For any model class $\mathscr{M} \subseteq \mathscr{V}$ with $\dim\langle\mathscr{M}\rangle < \infty$, any weight function w and all $\delta \in (0,1)$ there exists $C > 0$ such that

$$\|u - u_{\mathscr{M},n}\| \leq \left(1 + \tfrac{2}{\sqrt{1-\delta}}\right)\|u - u_{\mathscr{M}}\|_{w,\infty}$$

holds with probability $1 - C\exp\left(-\frac{n}{2}\left(\frac{\delta}{K(U(\{u_{\mathscr{M}}\}-\mathscr{M}))}\right)^2\right)$. The constant C is independent of n and depends only polynomially on δ and $K(U(\{u_{\mathscr{M}}\} - \mathscr{M}))^{-1}$.

This result is a combination of Theorems 2.8 and 2.12. Their proofs as well as the definitions of K, U and $\|\bullet\|_{w,\infty}$ can be found in Section 2. Some special model classes for which the theorem holds are discussed in Section 3.

To prove this result for general nonlinear model classes, we extend the idea of a restricted isometry property (RIP) as known from compressed sensing. In contrast to previous specific results for linear spaces [14], sets of sparse functions [38, 39], and low-rank tensors [40], the aim of this paper is to develop first results for a more general theory. New results for low-rank tensors are obtained and it is demonstrated how the theory can guide the choice of the model class $\mathscr{M}$.

Despite the generality of the derived theory we observe many of the same phenomena as more specialised theories namely, the emergence of an optimal sampling measure (*cf.* [14]), the importance of weighted sparsity (*cf.* [39]) and the advantage of multilevel sampling (*cf.* [3]).

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 and introduce the notion of the restricted isometry property (RIP). Based on 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. Section 3.2 considers sets of sparse functions and Section 3.3 examines sets of low-rank functions. Finally, we investigate the influence of the seminorm on the convergence 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

When $|v|_y = |v(y)|$ is used, $u_{\mathscr{M},n}$ is known as the nonlinear least squares estimator of u. The extensive interest in machine learning in recent years has lead to the investigation of this estimator for special model classes like sparse vectors [11,21,39], low-rank tensors [10,20,25,40,45] and neural networks [5,31]. However, to the knowledge of the authors no investigation for general model classes has been published so far. This may be due to the fact that sparse vectors and low-rank tensors were the first model classes for which rigorous theories were developed and that most of these works focus on ℓ^1 and nuclear norm minimization. Our work may be regarded as an extension of these works (in particular of infinite-dimensional compressed sensing [1,3]) to the nonlinear least-squares setting. For a more in-depth discussion of statistical learning theory we refer to the articles [15,43] and the monographs [16,26]. For linear spaces the first estimate in Theorem 2.12 has already appeared in [14] for weighted least squares and in [13,33,34] for standard least squares.

A convergence bound for the nonlinear least squares approximation problem was recently analysed in [20]. However, the probability of the bound failing increases exponentially as the best approximation error $\|u - u_{\mathscr{M}}\|$ approaches zero and becomes one when $\|u - u_{\mathscr{M}}\|$ vanishes. Moreover, this bound only holds for model classes that are bounded in L^∞ and it does not provide any insight on what property of the set influences the convergence rate.

The empirical approximation problem (1.1) was thoroughly examined in [14] for linear model spaces. There the model class $\mathscr{M}$ is assumed to be the m-dimensional subspace spanned by the *orthonormal* basis functions $\{\mathbf{B}_j\}_{j\in[m]}$ in $\mathscr{V} = L^2(Y,\rho)$. A key point in this work is that the error $\|u - u_{\mathscr{M},n}\|$ can be bounded by $\|u - u_{\mathscr{M}}\|_{L^\infty(Y,\rho)}$ if $\|\mathbf{G} - \mathbf{I}_m\|_2 \le \delta < 1$ where

$$\begin{aligned}\mathbf{G} &:= \frac{1}{n}\sum_{i=1}^n w(y_i)\mathbf{B}(y_i)\mathbf{B}(y_i)^\intercal \\ &= \frac{1}{n}\sum_{i=1}^n w(y_i)[\mathbf{B}_1(y_i)\ \dots\ \mathbf{B}_m(y_i)]^\intercal[\mathbf{B}_1(y_i)\ \dots\ \mathbf{B}_m(y_i)]\end{aligned}$$

is the Monte Carlo estimate of the Gram matrix $\mathbf{I}_m$. This condition is in fact equivalent to the norm equivalence

$$(1-\delta)\|u\|^2 \le \|u\|_n^2 \le (1+\delta)\|u\|^2 \qquad \text{for } u \in \mathscr{M}. \tag{1.2}$$

Cohen and Migliorati [14] prove that, under suitable conditions, the norm equivalence (1.2) is satisfied with high probability.

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

$$\mathbb{P}[\|\mathbf{G} - \mathbf{I}_m\|_2 > \delta] \le 2m\exp\left(-\frac{c_\delta n}{\tilde{K}}\right),$$

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

Equation (1.2) can be seen as a generalized *restricted isometry property*. The notion of a RIP was introduced in the context of compressed sensing [11]. It expresses the well-posedness of the problem by ensuring that $\|\bullet\|_n$ is indeed a norm and equivalent to $\|\bullet\|$ on $\mathscr{M}$. Minimizing the error with respect to $\|\bullet\|_n$ thus minimizes the error with respect to $\|\bullet\|$. In compressed sensing of sparse vectors [11,21] and low-rank tensors [40] discrete analogues of (1.2) are employed to derive bounds for the corresponding reconstruction errors. A recent work [6] also generalizes the theory from [14] to sparse grid spaces.

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

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

We prove that, under some conditions on n and A, this RIP holds with high probability and show that these conditions are satisfied for a variety of model classes. We then use the RIP to provide quasi-optimality guarantees for the empirical best approximation in Theorem 2.12.

In Remark 2.5 we note that it suffices to consider conic model sets. Optimizing over these sets is not straightforward. In [41], appropriate restricted isometry constants for exact recovery of conic model sets using a suitable regularizer are derived.

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 on the choice of w, $|\bullet|_y$ and Y. When a distinction between different choices of these parameters is necessary we add appropriate subscripts to K.

Remark 2.1. The variation constant $K(U(A))$ can be seen as a generalization of the embedding constant $(A, \|\bullet\|) \hookrightarrow \left(A, \|\bullet\|_{w,\infty}\right)$ to nonlinear sets and therefore as an analog of $\tilde{K}$ in Theorem 1.1.

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 $A \subseteq \mathscr{V}$ is defined by

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

This error is closely related to the RIP through the *normalization operator* U. This relation is developed rigorously in the subsequent lemma.

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

$$U(A) := \left\{\tfrac{u}{\|u\|} : u \in A\backslash\{0\}\right\}.$$

Lemma 2.4 (Equivalence of RIP and a bounded quadrature error)**.** *For some set A,*

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

Proof. Note that $\|0\|_n = \|0\|$, that $\|\alpha u\|_n = |\alpha|\|u\|_n$ for all $\alpha \in \mathbb{R}$ and $u \in A$ and that $\|u\| = 1$ for all $u \in U(A)$. Therefore,

$$\begin{aligned}
(1-\delta)\|u\|^2 \leq \quad & \|u\|_n^2 \quad \leq (1+\delta)\|u\|^2 && \forall u \in A\\
\Leftrightarrow \qquad (1-\delta) \leq \quad & \left\|\tfrac{u}{\|u\|}\right\|_n^2 \leq (1+\delta) && \forall u \in A\setminus\{0\}\\
\Leftrightarrow \qquad -\delta \leq \|u\|_n^2 - \|u\|^2 & \leq \delta && \forall u \in U(A),
\end{aligned}$$

which is equivalent to $\sup_{u\in U(A)}\left|\|u\|^2 - \|u\|_n^2\right| \leq \delta$. □

Remark 2.5. By the preceding lemma

$$\mathrm{RIP}_A(\delta) \Leftrightarrow \mathrm{RIP}_{\mathrm{Cone}(A)}(\delta),$$

where $\mathrm{Cone}(A) := \{\alpha a : a \in A, \alpha > 0\}$ denotes the cone generated by A. This implies that our theory also holds for unbounded sets 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 $A \subseteq \mathscr{V}$ is the minimal number of $\|\bullet\|$-open balls of radius ε that are needed to cover A.

Lemma 2.7. *Let $A \subseteq \mathscr{V}$ and $K = K(U(A)) < \infty$. Then,*

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

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 $A \subseteq \mathscr{V}$ and $K = K(U(A)) < \infty$. Then,*

$$\mathbb{P}[\mathrm{RIP}_A(\delta)] \geq 1 - 2\nu_{\|\bullet\|_{w,\infty}}\Big(U(A), \tfrac{1}{8}\tfrac{\delta}{\sqrt{K}}\Big)\exp\Big(-\tfrac{n}{2}\big(\tfrac{\delta}{K}\big)^2\Big) \quad \text{for} \quad \delta > 0.$$

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 complementary event. □

Corollary 2.9 (Sample complexity)**.** *Let $c, C, M > 0$ and $A \subseteq \mathscr{V}$ be a set that satisfies $\nu_{\|\bullet\|_{w,\infty}}(U(A), r) \leq C(cr)^{-M}$. Under the assumptions of Theorem 2.8, and with $K = K(U(A))$,*

$$n \geq 2\left(M\ln\left(\frac{8\sqrt{K}}{c\delta}\right) - \ln\Big(\frac{p}{2C}\Big)\right)\left(\frac{K}{\delta}\right)^2$$

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

Proof. To obtain $\mathrm{RIP}_A(\delta)$ with a probability of $1-p$ it suffices that

$$\mathbb{P}[\mathrm{RIP}_A(\delta)] \geq 1 - 2\nu\exp\Big(-\tfrac{n}{2}\big(\tfrac{\delta}{K}\big)^2\Big) \geq 1 - p,$$

with $\nu := \nu_{\|\bullet\|_{w,\infty}}\Big(U(A), \frac{1}{8}\frac{\delta}{\sqrt{K}}\Big)$. Solving the second inequality for n shows that

$$n \geq 2\ln\Big(\tfrac{2\nu}{p}\Big)\big(\tfrac{K}{\delta}\big)^2$$

samples are sufficient to satisfy $\mathrm{RIP}_A(\delta)$ with the prescribed probability. Replacing ν by the upper bound $C\Big(\frac{c\delta}{8\sqrt{K}}\Big)^{-M}$ yields the claim. □

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 [5,40,44]. The corollary states that in these cases $n \in \mathcal{O}(MG)$ where the factor $G := \ln(K)K^2$ represents the variation of $\|\bullet\|_n$ on $\mathscr{M}$.

Remark 2.10. An interpretation of Corollary 2.9 is 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 is independent of the dimension)**.** If $\mathscr{M}$ is a manifold then one might expect the bound for the probability of $\mathrm{RIP}_{\mathscr{M}}(\delta)$ to depend on its dimension. But counter-examples can be constructed easily. Consider $\mathscr{V} = L^2\big([-1,1], \frac{\mathrm{d}x}{2}\big)$ with the weight function $w \equiv 1$ and let P_k denote the k-th Legendre polynomial. Let moreover $d \in \mathbb{N}$ and $m \geq \frac{d^2}{2} + d - \frac{1}{2}$. Then the 1-dimensional manifold $\mathrm{span}\{P_m\}$ has a larger variation constant than the d-dimensional manifold $\mathrm{span}\{P_k\}_{k\in[d]}$. We refer to Section 3.1 for the computation of these variation constants. In the light of Remark 2.10, this means that the dimension d only has a minor influence on the bound for the probability of $\mathrm{RIP}_{\mathscr{M}}(\delta)$.

Theorem 2.12 (Empirical projection error)**.** *Assume that* $\mathrm{RIP}_{\{u_{\mathscr{M}}\}-\mathscr{M}}(\delta)$ *holds. Then*

$$\|u_{\mathscr{M}} - u_{\mathscr{M},n}\| \leq 2\frac{1}{\sqrt{1-\delta}}\|u - u_{\mathscr{M}}\|_{w,\infty}. \tag{2.1}$$

If in addition $\mathrm{RIP}_{\{u-u_{\mathscr{M}}\}}(\delta)$ *is satisfied then*

$$\|u_{\mathscr{M}} - u_{\mathscr{M},n}\| \leq 2\sqrt{\frac{1+\delta}{1-\delta}}\|u - u_{\mathscr{M}}\| \tag{2.2}$$

and consequently

$$\|u - u_{\mathscr{M}}\| \leq \|u - u_{\mathscr{M},n}\| \leq \left(1 + 2\frac{\sqrt{1+\delta}}{\sqrt{1-\delta}}\right)\|u - u_{\mathscr{M}}\|. \tag{2.3}$$

Proof. First observe that $u_{\mathscr{M},n} \in \mathscr{M}$ and therefore $u_{\mathscr{M}} - u_{\mathscr{M},n} \in \{u_{\mathscr{M}}\} - \mathscr{M}$. By the $\mathrm{RIP}_{\{u_{\mathscr{M}}\}-\mathscr{M}}(\delta)$, the triangle inequality and the definition of $u_{\mathscr{M},n}$, we deduce

$$\begin{aligned}\|u_{\mathscr{M}} - u_{\mathscr{M},n}\| &\leq \frac{1}{\sqrt{1-\delta}}\|u_{\mathscr{M}} - u_{\mathscr{M},n}\|_n \\ &\leq \frac{1}{\sqrt{1-\delta}}\left[\|u_{\mathscr{M}} - u\|_n + \|u - u_{\mathscr{M},n}\|_n\right] \\ &\leq 2\frac{1}{\sqrt{1-\delta}}\|u - u_{\mathscr{M}}\|_n.\end{aligned}$$

Hence, equation (2.1) holds since $\|v\|_n \leq \|v\|_{w,\infty}$ is satisfied for all $v \in \mathscr{V}$ and in particular for $u - u_{\mathscr{M}}$. Equation (2.2) follows by an application of $\mathrm{RIP}_{\{u-u_{\mathscr{M}}\}}(\delta)$ and from it equation (2.3) follows by an application of the triangle inequality to $\|u - u_{\mathscr{M},n}\|$. □

Remark 2.13. Note that Theorem 2.12 bounds $\|u_{\mathscr{M}} - u_{\mathscr{M},n}\|$ even if $u_{\mathscr{M}}$ and $u_{\mathscr{M},n}$ are not uniquely defined.

Remark 2.14. Theorem 2.12 requires $\mathrm{RIP}_{\{u_{\mathscr{M}}\}-\mathscr{M}}(\delta)$ and $\mathrm{RIP}_{\{u-u_{\mathscr{M}}\}}(\delta)$. If the covering number of $U(\{u_{\mathscr{M}}\} - \mathscr{M})$ is finite then $K(U(\{u_{\mathscr{M}}\} - \mathscr{M}))$ and $K(U(\{u - u_{\mathscr{M}}\}))$ are bounded and Theorem 2.8 guarantees that $\mathrm{RIP}_{\{u_{\mathscr{M}}\}-\mathscr{M}}(\delta)$ and $\mathrm{RIP}_{\{u-u_{\mathscr{M}}\}}(\delta)$ hold when n is sufficiently large.

If $u \in \mathscr{M}$ then $\mathrm{RIP}_{\{u-u_{\mathscr{M}}\}}(\delta)$ is implied by $\mathrm{RIP}_{\{u_{\mathscr{M}}\}-\mathscr{M}}(\delta)$ and bounds for the sample complexity of some well-known model classes are given in Section 3. If $u \notin \mathscr{M}$ then the probability of $\mathrm{RIP}_{\{u-u_{\mathscr{M}}\}}(\delta)$ has to be bounded separately. Since $\nu(U(\{u - u_{\mathscr{M}}\}), r) = 1$, we only need to bound $K(U(\{u - u_{\mathscr{M}}\}))$ to apply Theorem 2.8. Since $K(U(\{u - u_{\mathscr{M}}\}))$ depends only on $u - u_{\mathscr{M}}$ it is a purely approximation theoretic constant and we provide explicit bounds for two examples in the following.

- Let $\mathscr{M}$ be a space of low-rank functions and $\mathscr{M}_{\omega,s}$ be some space of sparse functions as defined in Section 3.2. If $u = u_{\text{low-rank}} + u_{\text{sparse}}$ with $u_{\text{low-rank}} \in \mathscr{M}$ and $u_{\text{sparse}} \in \mathscr{M}_{\omega,s}$ then $K(U(\{u - u_{\mathscr{M}}\})) \leq s^2$.
- Consider $u(x) := \sin(\pi x)$ and assume that $u_{\mathscr{M}}(x) = \sum_{k=0}^{m}(-1)^k \frac{(\pi x)^{2k+1}}{(2k+1)!}$ (for an otherwise arbitrary choice of $\mathscr{M}$). From this one can derive that $K(U(\{u - u_{\mathscr{M}}\})) \leq 4m + 7$.

Remark 2.15 (An indicator for $\mathrm{RIP}_A(\delta)$)**.** In Theorem 2.12 there is no constraint on the samples $\{y_i\}_{i=1}^n$ except that they satisfy the RIP. They explicitly do not have to be i.i.d. random variables. This means that they could, theoretically, be determined by a deterministic quadrature rule. The challenge, however, is to ensure the RIP. In [14] the empirical Gramian could be used to verify this RIP for a given sample set. In the nonlinear setting this is not possible. To obtain a practical indicator for the convergence of our method we make the following considerations. Define $A := (\{u_{\mathscr{M}}\} - \mathscr{M}) \cup \{u - u_{\mathscr{M}}\}$, as well as $e_n := \|u - u_{\mathscr{M},n}\|$ and $e := \|u - u_{\mathscr{M}}\|$. 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,

$$\mathbb{P}[\mathrm{RIP}_A(\delta)] \leq \mathbb{P}[e_n \leq (3 + 4\delta)e]. \tag{2.4}$$

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

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

Combining this with (2.4) yields

$$1 - \nu(\delta)\exp\left(-cn\delta^2\right) \leq \mathbb{P}[e_n \leq (3 + 4\delta)e] =: p(\delta). \tag{2.5}$$

Since $p(\delta)$ is increasing in δ, we can define an inverse in the sense of the quantile function $\delta(\tilde{p}) := \inf\left\{\tilde{\delta} \in \mathbb{R}_{\geq 0} : \tilde{p} \leq p\left(\tilde{\delta}\right)\right\}$. For fixed $\tilde{p} := p(\delta)$ in equation (2.5) it then follows that $\delta \geq \delta(\tilde{p}) =: \tilde{\delta}$ and consequently

$$-\ln(1 - \tilde{p}) \geq cn\tilde{\delta}^2 - \ln\left(\nu\left(\tilde{\delta}\right)\right)$$

or equivalently

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

Since $\delta(p) \geq 0$ is increasing and $-\ln(\nu(\delta(p))) \xrightarrow{p\to 1} 0$, the second term in the above sum becomes negligible for large p. This yields $\delta(p) \lesssim n^{-1/2}$, from which follows that $e_n \lesssim \left(1 + n^{-1/2}\right)e$. Hence, if $\mathrm{RIP}_A(\delta)$ holds for some $\delta \leq \frac{1}{\sqrt{2}}$, one can expect a rate of convergence that is reminiscent of the convergence rates for classical Monte Carlo quadrature. We can use this observation as an indicator for when $\mathrm{RIP}_A(\delta)$ is attained. To do this we select a test set of n' samples and observe the test set error $\tilde{e}_n := \|u - u_{\mathscr{M},n}\|_{n'}$ as the number of samples n is increased. When $\tilde{e}_n$ begins to decrease with an algebraic rate of $(1 + n^{-r})$ we take this as an indication that $\mathrm{RIP}_A(\delta)$ is satisfied and that additional sampling is unnecessary. This is illustrated in Figure 1.

Remark 2.16 (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

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

and the perturbed empirical best approximation

$$u_{\mathscr{M},n,\eta} \in \arg\min_{v\in\mathscr{M}} \|u - v\|_{\eta,n}.$$

Assume that $\mathrm{RIP}_{\{u_{\mathscr{M}}\}-\mathscr{M}}(\delta)$ holds. Then

$$\|u_{\mathscr{M}} - u_{\mathscr{M},n,\eta}\| \leq 2\frac{1}{\sqrt{1-\delta}}\|u - u_{\mathscr{M}}\|_{w,\infty} + \varepsilon.$$

If in addition $\mathrm{RIP}_{\{u-u_{\mathscr{M}}\}}(\delta)$ is satisfied then

$$\|u_{\mathscr{M}} - u_{\mathscr{M},n,\eta}\| \leq 2\sqrt{\frac{1+\delta}{1-\delta}}\|u - u_{\mathscr{M}}\| + \varepsilon.$$

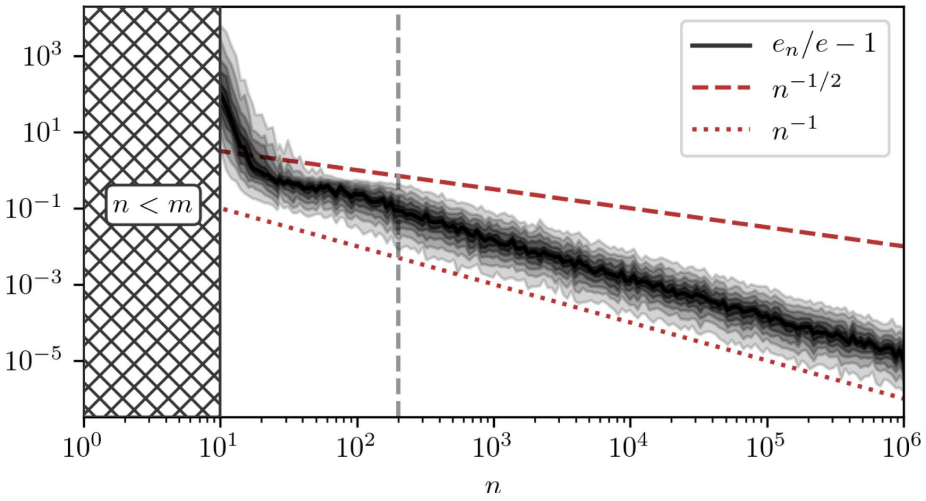

FIGURE 1. Let $\mathscr{V} = L^2\big([-1,1], \frac{\mathrm{d}x}{2}\big)$, $w \equiv 1$ and $\mathscr{M}$ be the model space of polynomials of degree less than $m = 10$. Let moreover A, e_n and e be defined as in Remark 2.15. Depicted is the distribution of the random variable $e_n/e - 1$ for different values of n and a synthetic (but fixed) function u. 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 $\mathrm{RIP}_A(\delta)$ w.r.t. n. From there on, $\mathrm{RIP}_A(\delta)$ holds with a high probability and the error decays with a rate of n^{-1}. Remark 2.15 predicts a rate of $n^{-1/2}$ but the condition $e_n \leq c(1 + n^{-r})$ is satisfied for $c = r = 1$. This faster decay can be explained by the fact that for the linear space $\mathscr{M}$ the bounds in the proof of Theorem 2.8 are suboptimal (see Example 3.1).

Remark 2.17. A generalization of Theorem 2.12 also holds for the residual minimization problem

$$v_{\mathscr{M}} \in \operatorname*{arg\,min}_{v \in \mathscr{M}} \|u - Lv\| \quad \text{and} \quad v_{\mathscr{M},n} \in \operatorname*{arg\,min}_{v \in \mathscr{M}} \|u - Lv\|_n,$$

since, whenever $\mathrm{RIP}_{\{u\}-L\mathscr{M}}(\delta)$ and $\mathrm{RIP}_{\{u - Lv_{\mathscr{M}}\}}(\delta)$ hold, we can estimate

$$\|u - Lv_{\mathscr{M},n}\| \lesssim \|u - Lv_{\mathscr{M},n}\|_n \leq \|u - Lv_{\mathscr{M}}\|_n \lesssim \|u - Lv_{\mathscr{M}}\|.$$

This means that the present theory can treat residual minimization problems by considering the RIP for the transformed model class $L\mathscr{M}$. An important application of such a problem arises in medical imaging and is briefly discussed in Example 4.3.

3. Examples and numerical illustrations

In this section, we examine some exemplary model classes 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 $\mathscr{V}$ be a separable vector space and $A \subseteq \mathscr{V}$. Then the pointwise supremum $\hat{b}(y) \coloneqq \sup_{v \in A} |v|_y^2$ with respect to $y \in Y$ is measurable and for any weight function w*

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

If A is $\|\bullet\|$-bounded, $K(A)$ is finite, and $\hat{b} > 0$ is almost surely positive, then

$$K(A) \geq \left\|\hat{b}\right\|_{L^1(Y,\rho)},$$

where the lower bound is attained by the weight function $w = \left\|\hat{b}\right\|_{L^1(Y,\rho)} \hat{b}^{-1}$.

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 spaces

Consider an m-dimensional linear subspace $\mathscr{V}_m \subseteq \mathscr{V} := L^2(Y,\rho)$ spanned by the *orthonormal* basis $\{\mathbf{B}_j\}_{j\in[m]}$. Recall that Theorem 3.1 implies $K(U(\mathscr{V}_m)) = \left\|w\hat{b}\right\|_{L^\infty(Y,\rho)}$ where

$$\hat{b}(y) = \sup_{\substack{v\in\mathscr{V}_m \\ \|v\|=1}} |v(y)|^2 = \sup_{\substack{\mathbf{v}\in\mathbb{R}^m \\ \|\mathbf{v}\|_2=1}} |\mathbf{B}(y)^\intercal \mathbf{v}|^2 = \|\mathbf{B}(y)\|_2^2.$$

Here, the second equality follows by orthonormality and the third by the Cauchy–Schwarz inequality. From this, Theorem 3.1 implies

$$K(U(\mathscr{V}_m)) \geq \left\|\hat{b}\right\|_{L^1(Y,\rho)} = m \tag{3.1}$$

where the optimal weight function is given by $w(y) \coloneqq m\|\boldsymbol{B}(y)\|_2^{-2}$. Note that this observation was already reported in [14].

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

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

Corollary 2.9 then bounds the sample complexity of this model class by

$$n \geq 2\left(m\ln\left(\frac{8m^{3/2}}{\delta}\right) - \ln\left(\frac{p}{2}\right)\right)\left(\frac{m}{\delta}\right)^2 \in \mathcal{O}\big(m^3\ln(m)\big),$$

when the optimal weight function is used. Although our approach is more general the resulting asymptotic bound differs only by a factor of m^2 from the bound $n \in \mathcal{O}(m\ln(m))$ provided in [14]. The near optimal bound in [14] is obtained by using tighter concentration inequalities (*cf.* [42]) when bounding the probability of $\mathrm{RIP}_{\mathscr{V}_m}(\delta)$ in Theorem 1.1.

Remark 3.2. When the sampling density cannot be changed, the variation constant can also be used to guide the choice of a suitable model class. For linear spaces this section shows that an optimal model space is spanned by an orthonormal basis for which the basis functions are bounded by 1. Such spaces are characterized in [30] and a prime example is the Fourier basis of $L^2\big([-1,1], \frac{\mathrm{d}x}{2}; \mathbb{C}\big)$.

3.2. Sets of sparse functions

In this section we follow the ideas of [39] 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 $\omega \preceq \tilde{\omega}$ (*i.e.* $\omega_j \leq \tilde{\omega}_j$ for all j) implies $\omega(S) \leq \tilde{\omega}(S)$ and that $\omega(S) = |S|$ for $\omega \equiv \mathbf{1}$.

Let in the following $\{\mathbf{B}_j\}_{j\in\mathbb{N}}$ be a fixed *orthonormal* basis for $\mathscr{V} := L^2(Y,\rho)$, fix a weight function w and define the model set

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

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

Lemma 3.3. *It holds that*

- $\mathscr{M}_{\tilde{\omega},s} \subseteq \mathscr{M}_{\omega,s}$ *for* $\omega \preceq \tilde{\omega}$,
- $\mathscr{M}_{\omega,s} \subseteq \mathscr{M}_{\omega,t}$ *for* $s \leq t$,
- $\mathscr{M}_{\omega,s} = -\mathscr{M}_{\omega,s}$ *and*
- $\mathscr{M}_{\omega,s} + \mathscr{M}_{\omega,t} \subseteq \mathscr{M}_{\omega,s+t}$.

Moreover, if $\omega_j \geq \|\mathbf{B}_j\|_{w,\infty}$ *for all* j *then* $\|v\|_{w,\infty} \leq \sqrt{s}\|v\|$ *for all* $v \in \mathscr{M}_{\omega,s}$.

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

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

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

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

Lemma 3.4. *Let* $c(\omega,w) := \sup_{j\in\mathbb{N}} \frac{\|\mathbf{B}_j\|_{w,\infty}}{\omega_j}$. *Then*

$$K(U(\mathscr{M}_{\omega,s})) \leq c(\omega,w)^2 s.$$

Proof. First, observe that $\mathscr{M}_{\omega,s} = \mathscr{M}_{c\omega,c^2 s}$ for any $c > 0$. Since $c(\omega,w)\omega_j \geq \|\mathbf{B}_j\|_{w,\infty}$ for all $j \in \mathbb{N}$, the claim follows directly from Lemma 3.3. $\square$

This setting also incorporates the standard sparsity class $\mathscr{M}_{\mathbf{1},k} \cap \mathscr{V}_m$ where $\mathbf{1} = (1,1,\ldots)$ and $\mathscr{V}_m$ is the m-dimensional space spanned by $\{\mathbf{B}_j\}_{j\in[m]}$. To see this, define for any arbitrary sequence $\omega \in \mathbb{R}^{\mathbb{N}}_{\geq 0}$ the restricted sequence $\omega^{\leq m}$ by

$$\omega_j^{\leq m} := \begin{cases} \omega_j & j \leq m \\ \infty & \text{otherwise} \end{cases}$$

and observe that $\mathscr{M}_{\mathbf{1},k} \cap \mathscr{V}_m = \mathscr{M}_{\mathbf{1}^{\leq m},k}$.

Remark 3.5. Consider the standard sparsity class $\mathscr{M}_{\mathbf{1},k} \cap \mathscr{V}_{d^M}$, where $\mathscr{V}_{d^M} = \tilde{\mathscr{V}}_d^{\otimes M}$ and $w = \tilde{w}^{\otimes M}$ both have product structure. If $\tilde{\mathscr{V}}_d$ is spanned by the basis $\{\tilde{\mathbf{B}}_j\}_{j\in[d]}$, as is common *e.g.* in polynomial regression, then

$$c\big(\mathbf{1}^{\leq d^M}, w\big) = \Big(\max_{j\in[d]} \big\|\tilde{\mathbf{B}}_j\big\|_{\tilde{w},\infty}\Big)^M.$$

This means that $K(U(\mathscr{M}_{\mathbf{1},k} \cap \mathscr{V}_{d^M}))$ grows exponentially with M, which limits the applicability of classical isotropic sparsity in high-dimensional settings.

Lemma 3.6. *Let $\mathscr{V}_m$ be the m-dimensional subspace spanned by $\{\mathbf{B}_j\}_{j\in[m]}$. Then there exists $C > 0$ such that*

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

Proof. Since $\mathscr{M}_{\omega,s} = \mathscr{M}_{c\omega,c^2 s}$ for any $c > 0$, it suffices to consider the case $c\big(\omega^{\leq m}, w\big) = 1$, *i.e.* $\omega_j \geq \|\mathbf{B}_j\|_{w,\infty}$ for all $j \in [m]$. In this case, we show that

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

For the first step, let $\{v_j\}$ be the centers of a $\|\bullet\|$-covering of $U(\mathscr{M}_{\omega,s} \cap \mathscr{V}_m)$ with radius $\frac{r}{\sqrt{2s}}$. Thus, for any $v \in U(\mathscr{M}_{\omega,s} \cap \mathscr{V}_m)$ there exists v_j such that $\|v - v_j\| \leq \frac{r}{\sqrt{2s}}$. Since $v - v_j \in \mathscr{M}_{\omega,2s}$ and by Lemma 3.3,

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

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

For the second step, observe that $\mathscr{M}_{\omega,s} \subseteq \mathscr{M}_{\mathbf{1},s} = \mathscr{M}_{\mathbf{1},\lfloor s\rfloor}$. Since $(\mathscr{V}_m, \|\bullet\|) \simeq (\mathbb{R}^m, \|\bullet\|_2)$ 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 [44] by

$$\nu_{\|\bullet\|_2}\left(S_1^{\mathbb{R}^m}(0) \cap \mathscr{M}_{\mathbf{1},\lfloor s\rfloor}, \frac{r}{\sqrt{2s}}\right) \leq \left(\frac{Cm\sqrt{2s}}{r\lfloor s\rfloor}\right)^{\lfloor s\rfloor} \leq \left(\frac{4Cm}{r\sqrt{s}}\right)^s.$$

This proves the claim. □

Theorem 3.7. *Let $\mathscr{V}_m$ be the m-dimensional subspace spanned by $\{\mathbf{B}_j\}_{j\in[m]}$. Then there exists $C > 0$ such that*

$$\mathbb{P}\big[\mathrm{RIP}_{\mathscr{M}_{\omega,s}\cap\mathscr{V}_m}(\delta)\big] \geq 1 - 2\exp\left(c\big(\omega^{\leq m}, w\big)^2 s \ln\big(\tfrac{8Cm}{\delta}\big) - \tfrac{n}{2}\Big(\tfrac{\delta}{c(\omega^{\leq m},w)^2 s}\Big)^2\right).$$

Proof. The assertion follows directly from Theorem 2.8 together with Lemmas 3.4 and 3.6. □

Remark 3.8. Assuming $\omega_j \geq \|\mathbf{B}_j\|_{w,\infty}$ for all $j \in [m]$, Theorem 3.7 implies that

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

samples are sufficient to satisfy $\mathrm{RIP}_{\mathscr{M}_{\omega,s}\cap\mathscr{V}_m}(\delta)$ with probability $1-p$. This result can be compared with Theorem 5.2 in [39] where

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

or Theorems 4.4 and 8.4 in [38] where

$$n \gtrsim s\max\big\{\ln^2(s)\ln(m)\ln(n), \ln\big(p^{-1}\big)\big\}\delta^{-2}c\big(\mathbf{1}^{\leq m}, w\big)^2.$$

Since our theory is very general, we cannot expect our bound to be as strong as these specialized bounds. This comparison, however, shows that our bound remains qualitatively similar up to polynomial factors.

Example 3.9. Consider the basis of tensorized Legendre polynomials $\mathbf{B}_\mathbf{j} = \bigotimes_{m=1}^M \mathbf{L}_{\mathbf{j}_m}$ and define the linear space $\mathscr{V}_m$ in Theorem 3.7 as $\mathscr{V}_m := \operatorname{span}\left\{\mathbf{B}_\mathbf{j} : \omega_\mathbf{j}^2 \leq s\right\}$. Then the bound in Theorem 3.7 depends on the parameter s alone since the size of the hyperbolic cross

$$\left\{\mathbf{j} \in \mathbb{N}^M : \omega_\mathbf{j}^2 \leq s\right\} \subseteq \left\{\mathbf{j} \in \mathbb{N}^M : \prod_{m=1}^M (2\mathbf{j}_m + 1) \leq s\right\}$$

can be bounded by $m \lesssim s \log(s)^{M-1}$ (*cf.* [39]).

For a fixed basis $\{\mathbf{B}_j\}_{j\in\mathbb{N}}$ and weight sequence ω, the coefficient $c(\omega, w)$, defined in Lemma 3.4, depends only on the weight function w. Theorem 3.7 indicates that the probability of $\mathrm{RIP}_{\mathscr{M}_{\omega,s}\cap\mathscr{V}_m}(\delta)$ is maximized when

$$c\big(\omega^{\leq m}, w\big)^2 = \sup_{y\in Y} w(y) \max_{j\in[m]} \frac{|\mathbf{B}_j|_y^2}{\omega_j^2} =: \sup_{y\in Y} w(y)\tilde{b}(y) = \left\|w\tilde{b}\right\|_{L^\infty(Y,\rho)}$$

is minimized. From Theorem 3.1 we know that the minimum $\left\|w\tilde{b}\right\|_{L^\infty(Y,\rho)} = \left\|\tilde{b}\right\|_{L^1(Y,\rho)}$ is attained for the weight function $\tilde{w} = \left\|\tilde{b}\right\|_{L^1(Y,\rho)} \tilde{b}^{-1}$. Numerical experiments that compare different weight sequences ω and weight functions w are provided in Figure 2.

We finally note, that the theory presented in this subsection can be generalized easily to dictionary learning (*cf.* [19, 29]). This is stated, without proof, in the following theorem.

Theorem 3.10. *Assume that $\{\mathbf{B}_j\}_{j\in\mathbb{N}}$ is a Riesz sequence satisfying*

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

and that ω is chosen such that $\omega_j \geq \|\mathbf{B}_j\|$ for all j. Redefine

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

and let $\mathscr{V}_m \subset \mathscr{V}$ be the m-dimensional subspace spanned by $\{\mathbf{B}_j\}_{j\in[m]}$. Then it holds, that

- $\mathscr{M}_{\tilde{\omega},s} \subseteq \mathscr{M}_{\omega,s}$ *for* $\omega \preceq \tilde{\omega}$,
- $\mathscr{M}_{\omega,s} \subseteq \mathscr{M}_{\omega,t}$ *for* $s \leq t$,
- $\mathscr{M}_{\omega,s} = -\mathscr{M}_{\omega,s}$,
- $\mathscr{M}_{\omega,s} + \mathscr{M}_{\omega,t} \subseteq \mathscr{M}_{\omega,s+t}$,
- $\|v\|_{w,\infty} \leq \frac{\sqrt{s}}{c}\|v\|$ *for all* $v \in \mathscr{M}_{\omega,s}$ *and*
- $\nu_{\|\bullet\|_{w,\infty}}(U(\mathscr{M}_{\omega,s}\cap\mathscr{V}_m), r) \leq \left(k\frac{Cm}{cr\sqrt{s}}\right)^s$ *for some* $k > 0$.

3.3. Tensors of rank r

We now consider two different problems related to model classes of low-rank tensors. Both can be expressed with $Y = (\mathbb{R}^m)^{\otimes M}$ and $\mathscr{V} = (Y, \|\bullet\|)$ with $|v|_y := |(v, y)_{\mathrm{Fro}}|$. The only difference is the distribution ρ from which the samples are drawn.

(1) *Recovery from Gausian samples.* In this problem $\rho = \mathcal{N}(0_Y, \mathrm{Id}_Y)$ is a Gaussian distribution on the tensor space Y and $\|\bullet\| = \|\bullet\|_{\mathrm{Fro}}$. Although this problem is rather artificial it was one of the first where rigorous bounds were developed in [40].

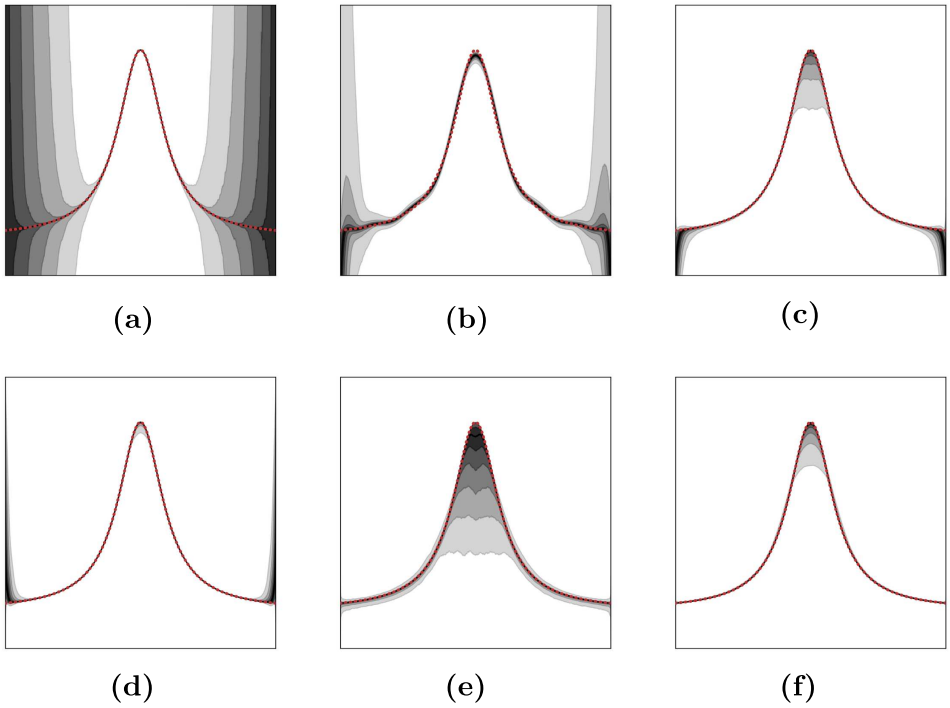

FIGURE 2. Interpolations of the function $f(x) = \frac{1}{1+25x^2}$ (red) by Legendre polynomials of degree 99, using six different methods. Depicted is the probability distribution of the function values for an interpolation that uses $n = 30$ random sampling points. The subfigures 2a and 2b show standard least squares approximations of the first 30 and 15 basis functions, respectively. The other figures employ the weighted ℓ^1-minimization $\min_{\mathbf{v}\in\mathbb{R}^{100}} \|\Omega\mathbf{v}\|_1$ s.t. $\mathbf{v}^\intercal \mathbf{B}(y_i) = f(x_i)$ for $1 \leq i \leq n$, with $\Omega := \operatorname{diag}(\omega)$. For 2c and 2d, the sampling points are drawn according to the uniform measure on $[-1, 1]$. For 2e and 2f, the sampling points are drawn according to the measure induced by $\tilde{b}$. Subplots 2c and 2e display the results of standard ℓ^1-minimization (*i.e.* $\omega_j = 1$) while 2d and 2f display the results of weighted ℓ^1-minimization (*i.e.* $\omega_j = \|\mathbf{B}_j\|_{L^\infty}$). (a) Exact inversion, mean error: 2×10^7. (b) Least squares, mean error: 3×10^2. (c) Standard ℓ^1, unweighted samples, mean error: 6×10^{-1}. (d) Weighted ℓ^1, unweighted samples, mean error: 5×10^{-1}. (e) Standard ℓ^1, weighted samples, mean error: 3×10^{-1}. (f) Weighted ℓ^1, weighted samples, mean error: 8×10^{-2}.

(2) *Recovery from rank-1 samples and completion.* For this problem let $\{\rho_k\}_{k\in[M]}$ be distributions on $\mathbb{R}^m$ and consider the push-forward of $\rho = \rho_1 \otimes \cdots \otimes \rho_M$ through the tensor product map. This means, that every realization $(y_1, \ldots, y_M) \sim \rho$ is mapped to the tensor product $y_1 \otimes \cdots \otimes y_M$. This problem occurs for example whenever one tries to approximate a low-rank function of M variables using a tensor product basis. A special case of this setting is the problem of tensor completion where a tensor has to be recovered from a few of its entries. In this case, all distributions ρ_k are discrete measures on the standard basis vectors.

In both problems the task is to find a best approximation in a subset $\mathscr{T}_r \subseteq \mathscr{V}$ of bounded rank r. For tensors, however, there exist many different concepts of rank for which we refer to [4, 24, 27, 28] and the works cited below.

Recovery from Gaussian samples

In this section we consider a subset $\mathscr{T}_r \subseteq \mathscr{V}$ of tensors of bounded (Hierarchical Tucker) HT-rank r. For $w \equiv 1$ the following bound for the sample complexity subject to δ is given in Theorem 2 of [40],

$$n \gtrsim \max\{\big((M-1)r^3 + Mmr\big)\ln(Mr), \ln\big(p^{-1}\big)\}\delta^{-2}.$$

To obtain a sample bound from our theory, we would have to bound the variation constant, which however is infinity,

$$K(U(\mathscr{T}_r)) = \sup_{\substack{v\in\mathscr{T}_r\\ \|v\|=1}} \operatorname*{ess\,sup}_{y\in Y} |(v,y)_{\mathrm{Fro}}|^2 = \infty.$$

This shows that a direct application of the presented formalism to this problem cannot provide a finite sample complexity.

Remark 3.11. As above, this exposes the lacking sharpness of the results used in the proof of Theorem 2.8. With more refined concentration inequalities as in [18], a different definition of the variation constant would emerge (replacing $\|\bullet\|_{w,\infty}$ by a sub-Gaussian norm), which would be finite for this problem.

The present theory can deal with this problem in two different ways. The first option is to choose the weight function $w(y) = m^M \|y\|_{\mathrm{Fro}}^{-2}$, which yields the variation constant

$$K(U(\mathscr{T}_r)) = m^M \sup_{\substack{v\in\mathscr{T}_r\\ \|v\|=1}} \operatorname*{ess\,sup}_{\substack{y\in Y\\ \|y\|=1}} |(v,y)_{\mathrm{Fro}}|^2 = m^M,$$

where the final equality holds since $\|\bullet\| = \|\bullet\|_{\mathrm{Fro}}$. The second option is to normalize the samples and thereby replace the Gaussian distribution by a uniform distribution on the unit sphere. In this case we obtain the new identity $\|\bullet\| = m^{-M/2}\|\bullet\|_{\mathrm{Fro}}$ and the corresponding variation constant

$$K(U(\mathscr{T}_r)) = m^M \sup_{\substack{v\in\mathscr{T}_r\\ \|v\|=1}} \operatorname*{ess\,sup}_{\substack{y\in Y\\ \|y\|=1}} |(v,y)_{\mathrm{Fro}}|^2 = m^M.$$

In both cases $K(U(\mathscr{T}_r)) = K(U(\mathscr{V}))$. Let $k = \sqrt{K(U(\mathscr{T}_{2r}))}$. By using the bound $\|\bullet\|_{w,\infty} \leq k\|\bullet\|$ on $\mathscr{T}_{2r}$ we can utilize the bound for the covering number for tensors of HT-rank r that is provided in [40]. This leads to the estimate

$$\nu_{\|\bullet\|_{w,\infty}}(U(\mathscr{T}_r),\varepsilon) \leq \nu_{\|\bullet\|}\Big(U(\mathscr{T}_r), \frac{\varepsilon}{k}\Big) \leq \left(\frac{\varepsilon}{3(2M-1)\sqrt{r}k}\right)^{-\left(Mr^3+Mmr\right)}.$$

A subsequent application of Corollary 2.9 yields

$$n \geq 2\Big(\big(Mr^3+Mmr\big)\ln\big(3(2M-1)\sqrt{r}k\delta^{-1}\big) - \ln\Big(\frac{p}{2}\Big)\Big)\left(\frac{k^2}{\delta}\right)^2.$$

For $k = 1$ this would have the same asymptotic complexity as the bound in [40] and we conjecture that the transition $k = m^{M/2} \rightsquigarrow k = 1$ can be achieved by using a generic chaining argument (*cf.* [18]) rather than a simple Hoeffding bound in the proof of Theorem 2.8.

*Recovery from rank-*1 *samples and completion*

In this section we consider subsets $\mathscr{T}_r \subseteq \mathscr{V}$ of rank-r tensors for any rank concept that satisfies $\mathscr{T}_1 \subseteq \mathscr{T}_r$. This is the case for all tree-shaped tensor formats including the Tucker format, the tensor train (TT) format and general hierarchical tensor formats (HT) as well as the canonical polyadic decomposition (CP). For the sake of completeness we define

$$\mathscr{T}_1 := \{v \in \mathscr{V} : \mathbf{v} = \mathbf{v}_1 \otimes \cdots \otimes \mathbf{v}_M \text{ with } \mathbf{v}_1, \ldots, \mathbf{v}_M \in \mathbb{R}^m\}.$$

The variation constant for the set $\mathscr{T}_r$ is computed in the next theorem.

Theorem 3.12. *$K(U(\mathscr{T}_r)) = K(U(\mathscr{V}))$ for any weight function w.*

Proof. Observe that

$$\hat{b}_{U(\mathscr{V})}(y) = \sup_{\substack{v\in\mathscr{V}\\ \|v\|=1}} |(v,y)_{\mathrm{Fro}}|^2 = \left|\left(\tfrac{y}{\|y\|}, y\right)_{\mathrm{Fro}}\right|^2 = \frac{\|y\|_{\mathrm{Fro}}^4}{\|y\|^2}.$$

Moreover, if y is of rank 1, then

$$\hat{b}_{U(\mathscr{T}_1)}(y) = \sup_{\substack{v\in\mathscr{T}_1\\ \|v\|=1}} |(v,y)_{\mathrm{Fro}}|^2 = \left|\left(\tfrac{y}{\|y\|}, y\right)_{\mathrm{Fro}}\right|^2 = \frac{\|y\|_{\mathrm{Fro}}^4}{\|y\|^2}.$$

Since the measure ρ is supported only on the set $\mathscr{T}_1$, we deduce, using Theorem 3.1, that $K(U(\mathscr{T}_1)) = K(U(\mathscr{V}))$. This proves the assertion since $\mathscr{T}_1 \subseteq \mathscr{T}_r \subseteq (\mathbb{R}^m)^{\otimes M}$ implies $K(U(\mathscr{T}_1)) \leq K(U(\mathscr{T}_r)) \leq K(U(\mathscr{V}))$. □

The theorem states that tensor formats do **not** exhibit a smaller variation constant than the linear space they are embedded in. This result is surprising at first because tensor formats have a significantly smaller covering number than the full tensor space, *cf.* [40]. However, this is already indicated by the classical analysis of matrix completion from which it is known that the notion of incoherence is required in addition to a low-rank property.

Despite this unfavourable result, it is noteworthy that the present theory can be used in this setting. The bound $\|\bullet\|_{w,\infty} \leq \sqrt{K(U(\mathscr{T}_{2r}))}\|\bullet\|$ and the isometry $\|\bullet\| = \|\bullet\|_{\mathrm{Fro}}$ imply

$$\nu_{\|\bullet\|_{w,\infty}}(U(\mathscr{T}_r), \varepsilon) \leq \nu_{\|\bullet\|_{\mathrm{Fro}}}\left(U(\mathscr{T}_r), \frac{\varepsilon}{\sqrt{K(U(\mathscr{T}_{2r}))}}\right).$$

Assuming the weight function w is chosen optimally, we know from Theorem 3.12 and equation (3.1) that $K(U(\mathscr{T}_{2r})) = m^M$. We can now apply the bound for the covering number of tensors of HT-rank r from [40]. The resulting estimate reads

$$\nu_{\|\bullet\|_{w,\infty}}(U(\mathscr{T}_r), \varepsilon) \leq \left(\frac{\varepsilon}{3(2M-1)\sqrt{rm^M}}\right)^{-\left(Mr^3+Mmr\right)}.$$

A final application of Corollary 2.9 yields

$$n \geq 2\Big(\left(Mr^3+Mmr\right)\ln\Big(3(2M-1)\sqrt{rm^M}\delta^{-1}\Big) - \ln\Big(\frac{p}{2}\Big)\Big)\left(\frac{m^M}{\delta}\right)^2.$$

To the knowledge of the authors this is the first estimate of the number of samples that are necessary to satisfy $\mathrm{RIP}_{\mathscr{T}_r}(\delta)$ in this setting. Note that this is a worst-case estimate and that significantly less samples are needed in practice (*cf.* [20]).

In the following examples we discuss the application to two common classes of problems.

Example 3.13. In this example we consider the problem of recovering the low-rank coefficient tensor of a function from samples. Let π_m be a probability measure on Z_m and $\mathscr{W}_m \subseteq L^2(Z_m, \pi_m)$ be spanned by the d_m *orthonormal* basis functions $\{\mathbf{B}_{m,j}\}_{j\in[d_m]}$. Now define the product space $\mathscr{W} := \mathscr{W}_m^{\otimes M} \subseteq L^2(Z,\pi)$ with $Z := Z_m^M$ and $\pi := \pi_m^{\otimes M}$ and endow it with the seminorm $\dagger w\dagger_z := |w(z)|$. For the sake of simplicity we assume that the weight function is constant, which means that writing $w \in \mathscr{W}$ does not introduce a conflict of notation. This is the space in which the sought functions will live and it shall be approximated in the norm $\ddagger\bullet\ddagger := \|\bullet\|_{L^2(Z,\pi)}$.

As a model class consider the set $\mathcal{T}_r^{\mathscr{W}} \subseteq \mathscr{W}$ of functions with a coefficient tensor of rank r. Note that every $w \in \mathcal{T}_r^{\mathscr{W}}$ can be represented in the tensor product basis $\mathbf{B}_\mathbf{j}(z) := \prod_{k=1}^M \mathbf{B}_{m,\mathbf{j}_k}(z_k)$ as

$$w(z) = \sum_{\mathbf{j}\in\mathbb{N}^M} \mathbf{w}_\mathbf{j} \mathbf{B}_\mathbf{j}(z)$$

with a coefficient tensors $\mathbf{w} \in \mathcal{T}_r^{\mathscr{V}} := \mathcal{T}_r$.

To compute the variation constant of this model class, recall the definition of $\mathscr{V} = (Y, \|\bullet\|)$, $Y = (\mathbb{R}^m)^{\otimes M}$ and $|v|_y = |(v,y)_{\mathrm{Fro}}|$ from above. Note that each function $w \in \mathscr{W}$ corresponds uniquely to a coefficient tensor $\mathbf{w} \in \mathscr{V}$ and that the mapping $\mathbf{B} : Z \to Y$ given by $(\mathbf{B}(z))_\mathbf{j} := \mathbf{B}_\mathbf{j}(z)$ induces an isometry of seminorms

$$\dagger w \dagger_z = |w(z)| = |(\mathbf{w}, \mathbf{B}(z))_{\mathrm{Fro}}| = |\mathbf{w}|_{\mathbf{B}(z)}.$$

This means that, if we choose ρ as the pushforward measure $\rho := \mathbf{B}_* \pi$, the isometry of seminorms induces the isometry of the two norms

$$\|\mathbf{w}\| = \left(\int_Y |\mathbf{w}|_y^2 \,\mathrm{d}\rho(y)\right)^{1/2} = \left(\int_Z \dagger w \dagger_z^2 \,\mathrm{d}\pi(z)\right)^{1/2} = \ddagger w \ddagger$$

and

$$\|\mathbf{w}\|_{1,\infty} = \operatorname*{ess\,sup}_{y\in Y} |\mathbf{w}|_y = \operatorname*{ess\,sup}_{z\in Z} \dagger w \dagger_z =: \ddagger w \ddagger_{1,\infty}.$$

Together with Theorems 3.12 and 3.1 it follows that

$$K\big(U\big(\mathcal{T}_r^{\mathscr{W}}\big)\big) = K\big(U\big(\mathcal{T}_r^{\mathscr{V}}\big)\big) = K(U(\mathscr{V})) \geq m^M.$$

This shows that the variation constant for this problem grows exponentially with M.

Example 3.14. The problem of tensor completion can be considered as a special case of Example 3.13. In this setting $Z = [m]^M$ is the set of all multi-indices, $\pi = \mathcal{U}(Z)$ is a uniform distribution on Z and $\mathscr{W} = \left((\mathbb{R}^m)^{\otimes M}, \|\bullet\|_{\mathrm{Fro}}\right)$ is endowed with the semi-norm $\dagger w \dagger_z := m^M |w_z|$. Since this is a special case of Example 3.13 the model class of rank-r tensors $\mathcal{T}_r$ exhibits the same bound, namely $K(U(\mathcal{T}_r)) = K(U(\mathscr{W})) \geq m^M$.

These two examples show that $K(U(\mathcal{T}_r)) \geq m^M$ in important applications. To reduce the variation constant in these cases we can only intersect $\mathcal{T}_r$ with another model class $\mathscr{M}$ with low variation constant. The intersection then inherits the low covering number of $\mathcal{T}_r$ and the low variation constant of $\mathscr{M}$.

4. Dependence on the seminorm

Since the definition of the $\|\bullet\|$-norm is very general, our theory is not limited to the L^2-norm but extends to Sobolev or energy norms. It is therefore natural to ask how the choice of the semi-norm $|\bullet|_y$ influences the variation constant. In this section we investigate this influence using Sobolev norms as an example.

We will need the following generalization of reproducing kernel Hilbert spaces (GRKHS) as a tool for the analysis.

Definition 4.1 (Generalized Reproducing Kernel Hilbert Space)**.** Let $\mathscr{H} \subseteq \mathscr{V}$ and $\{L_y\}_{y\in Y} \subseteq \mathscr{L}(\mathscr{H}, \mathbb{R}^\ell)$ be a family of bounded linear operators. Then the pair $\left(\mathscr{H}, \{L_y\}_{y\in Y}\right)$ generalizes the concept of reproducing kernel Hilbert spaces.

If $\left(\mathscr{H}, \{L_y\}_{y\in Y}\right)$ forms a GRKHS and $|v|_y := \|L_y v\|_2$ then

$$\|v\|_{w,\infty} \leq \varkappa \|v\|_{\mathscr{H}} \qquad \text{and} \qquad K(U(A)) \leq \varkappa^2 \lambda^2,$$

for $v \in A \subseteq \mathscr{H}$ with $\varkappa := \sup_{y\in Y} \sqrt{w(y)} \|L_y\|_{\mathscr{L}(\mathscr{H},\mathbb{R}^l)}$ and $\lambda := \sup_{v\in A\setminus\{0\}} \frac{\|v\|_{\mathscr{H}}}{\|v\|}$. This allows to efficiently compute an upper bound for $K(U(A))$.

Remark 4.2. In this setting the application of Theorem 2.12 leads to

$$\|u - u_{\mathscr{M},n}\| \lesssim \|u - u_{\mathscr{M}}\|_{w,\infty} \leq \varkappa \|u - u_{\mathscr{M}}\|_{\mathscr{H}}$$

whenever $\mathrm{RIP}_{\{u_{\mathscr{M}}\}-\mathscr{M}}(\delta)$ holds.

In the following, we consider a linear model space $\mathscr{M} \subseteq \mathscr{H} := H^M(Y)$ with a Lipschitz domain $Y \subseteq \mathbb{R}^d$. For each $m \leq M - \frac{d}{2}$ we consider $\mathscr{V} := H^m(Y)$ with $|v|_y := \left\|L_y^m v\right\|_2$ and $L_y^m \in \mathscr{L}\left(\mathscr{H}, \mathbb{R}^\ell\right)$ defined such that $\|\bullet\| = \|\bullet\|_{H^m}$. This means, that we are searching the best approximation in the model space $\mathscr{M}$ with respect to the H^m-norm. To investigate the influence of m on the sample complexity, the upper bound $\varkappa_m^2 \lambda_m^2$ for $K(U(\mathscr{M}))$ depending on m has to be computed.

It is proven in Appendix C, that for $w \equiv 1$

$$\varkappa_m := \left(2\sqrt{\pi}\right)^{-d} \frac{\Gamma(M+1)\Gamma\left(M - m - \frac{d}{2}\right)}{\Gamma(M-m)}. \tag{4.1}$$

Since $\varkappa_m$ increases whith m, while λ_m decreases, both effects should be equilibrated by a proper choice of m. This is illustrated for two different model spaces $\mathscr{M}$ in Figure 3. The small effect of $\varkappa_m$ is due to the dimension $d = 1$ for which we can bound (4.1) by

$$\frac{(M+1)!}{2\sqrt{\pi}}(M-m)^{-1/2} < \varkappa_m < \frac{(M+1)!}{2\sqrt{\pi}}(M-m-1)^{-1/2}$$

via Gautschi's inequality ([35], Eq. (5.6.4)).

We conclude that for linear model spaces an approximation with respect to the H^m-norm for larger m requires less samples than an approximation with respect to the L^2-norm. For $m = 1$ this hypothesis is confirmed numerically in Figure 4. For an application in the setting of weighted sparsity we refer to the recent work [2]. Note that this does not have to be the case in general. If the model class contains only piecewise constant functions then information about the gradients is irrelevant. Such phenomena may also arise due to intricate properties of the model class and may only be observable by looking at the variation constant.

Also note that the minimization with respect to the H^m-norm does not necessarily require more computational effort than the minimization with respect to the L^2-norm. The values of both seminorms can be computed with a single evaluation of the Fourier transform $\hat{u} = \mathscr{F}u$ of u. A particularly important application of this setting is *Magnetic Resonance Imaging* (MRI). Recalling Remark 2.17, we describe this application in the following example.

Example 4.3 (MRI)**.** In Magnetic Resonance Imaging an image u is sampled *via* evaluations of its Fourier transform $\hat{u} = \mathscr{F}u$. This means that the samples $\{\hat{u}_i\}_{i\in[n]}$ satisfy $\hat{u}_i = \hat{u}(\omega_i)$ for samples of the angular frequency ω_i. The precise distribution of the samples ω_i is given by the problem and is not of particular interest in this example. Since u is an image of the human body, we can assume that it can be sparsely represented in a wavelet basis (*cf.* [9,37]). The MRI reconstruction problem can hence be written as

$$v_{\mathscr{M}} \in \operatorname*{arg\,min}_{v\in\mathscr{M}_{\mathbf{1},k}} \|\hat{u} - \mathscr{F}v\|,$$

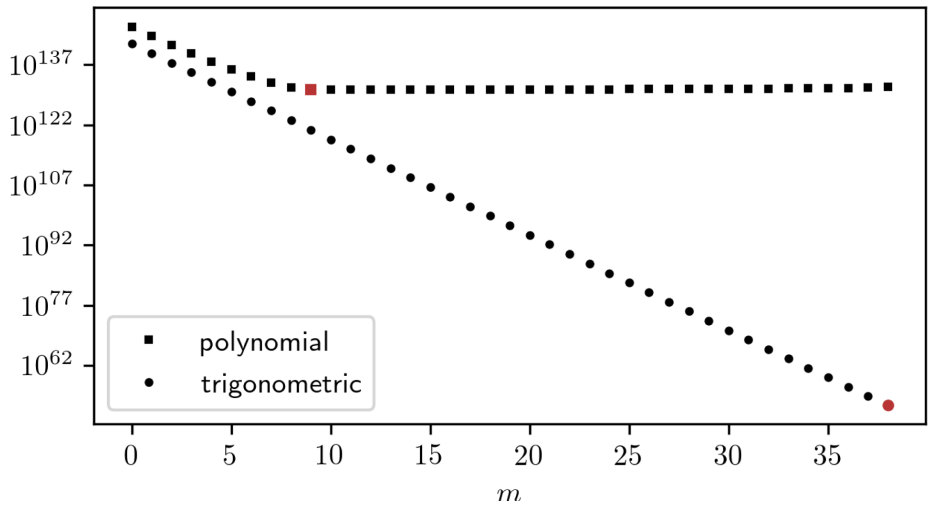

FIGURE 3. The upper bound $\varkappa_m^2 \lambda_m^2$ for the variation constant for two different model spaces $\mathscr{M}$ with $Y = [-1, 1]$ and $M = 40$. The squares and dots represent the bound when A is the span of the first 10 polynomials and trigonometric polynomials, respectively. The optimal m is marked fat and in red.

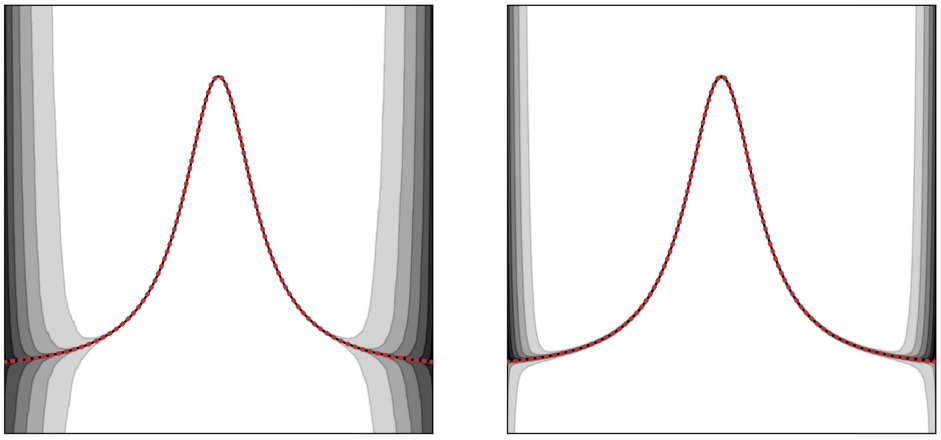

FIGURE 4. Probability distributions of the function values for least squares approximations of the function $f(x) = \frac{1}{1+25x^2}$ (red) by Legendre polynomials of degree 29. Different approximations correspond to different random draws of $n = 40$ sampling points from the uniform measure on $[-1, 1]$. (a) L^2-least squares, mean error: 2×10^7. (b) H^1-least squares, mean error: 5×10^5.

where the seminorm is chosen as $|v|_\omega = |v(\omega)|$ and $\mathscr{M}_{\mathbf{1},k}$ is defined with respect to the chosen wavelet basis. From Remark 2.17 we know that recovery requires $\mathrm{RIP}_{\{\hat{u}\}-\mathscr{F}\mathscr{M}_{\mathbf{1},k}}(\delta)$ and $\mathrm{RIP}_{\{\hat{u}-\mathscr{F}u_{\mathscr{M}}\}}(\delta)$, the probabilities of which can be bounded by Theorem 2.8.

In the following we only compute the variation constant since the Fourier transform is an isometry and does not change the covering number. Assuming that $u \in \mathscr{M}_{\mathbf{1},k}$, we can estimate

$$K(U(\{\hat{u} - \mathscr{F}v_{\mathscr{M}}\})) \leq K(U(\{\hat{u}\} - \mathscr{F}\mathscr{M}_{\mathbf{1},k})) \leq K(U(\mathscr{F}\mathscr{M}_{\mathbf{1},2k})).$$

To evaluate this, let ψ be the mother wavelet and define the daughter wavelets $\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\big(\frac{t-b}{a}\big)$. Due to basic properties of the Fourier transform $\hat{\psi}_{a,b}(\omega) := (\mathscr{F}\psi_{a,b})(\omega) = \sqrt{a}\hat{\psi}(a\omega)\exp(-ia\omega)$ and since the daughter wavelets are normalized we obtain

$$K\Big(U\Big(\Big\langle\hat{\psi}_{a,b}\Big\rangle\Big)\Big) = \frac{\Big\|\hat{\psi}_{a,b}\Big\|_{L^\infty}^2}{\Big\|\hat{\psi}_{a,b}\Big\|_{L^2}^2} = a\Big\|\hat{\psi}\Big\|_{L^\infty}^2.$$

Note that $\hat{\psi}$ is the Fourier transform of the mother wavelet and therefore $\Big\|\hat{\psi}\Big\|_{L^\infty}^2$ is constant. It can be concluded that many samples are needed to recover larger scale coefficients but fewer samples for smaller scales. This suggests a multilevel approach where the small-scale coefficients are learned separately from the large-scale coefficients. This was already observed in the compressed sensing literature (*cf.* [3]). Typically, these schemes use the classical unweighted notion of sparsity. For a recent application of weighted sparsity in the context of residual minimization in a sparse wavelet representation we refer to [17].

Due to the high variation constant of the large scale coefficients, it is sensible to incorporate as much information as possible into this model class. In the spirit of works like [12], this can for example be achieved by means of manifold constraints. These manifolds can either be estimated for a single patient (*cf.* [32]) or for multiple patients when it can be assumed that the large-scale structures remain similar for different patients. In this way the image u is decomposed (approximately) as a sum of a background image modelling the healthy tissue and a foreground image modelling the pathological lesion.

Note that, if the mother wavelet ψ is differentiable, we can also consider the semi-norm $|v|_\omega := \sqrt{1+\omega^2}|v(\omega)|$, which corresponds to the H^1-norm in the physical domain. Computing the variation constant is however out of scope for this brief discussion.

5. Discussion

The nonlinear least squares method is probably the easiest and, currently, the most commonly used setting in machine learning regression. In Section 2 we derive an error bound for the nonlinear least squares estimator (1.1) that can be used with arbitrary model classes. This result is based on a *restricted isometry property (RIP)*, which we prove to hold with high probability when the number of samples is sufficiently large.

To put our theory into perspective, we apply it to well-known model classes and compare the results to the near optimal bounds that often already exist in the literature. In the cases of linear spaces (Sect. 3.1), functions with sparse representation (Sect. 3.2) and low-rank tensors (Sect. 3.3), we obtain asymptotic bounds which differ from these near optimal ones by a polynomial factor. This means that our analysis does not provide optimal complexity bounds when the number of samples should be determined *a priori* and when sampling is costly (*i.e.* when it is imperative to require as few samples as possible). We however assume that a more meticulous application of modern concentration arguments (like [18]) could close this gap. We also obtain first bounds for the sample complexity of low-rank tensor recovery in Section 3.3. These bounds, however, only improve the sample complexity by a logarithmic term in comparison to full-rank tensors. An intuition for this result is already provided by matrix recovery, where it is known that regularity in the form of incoherence is needed in addition to the low-rank property. As a first remedy, we suggest to impose additional regularity

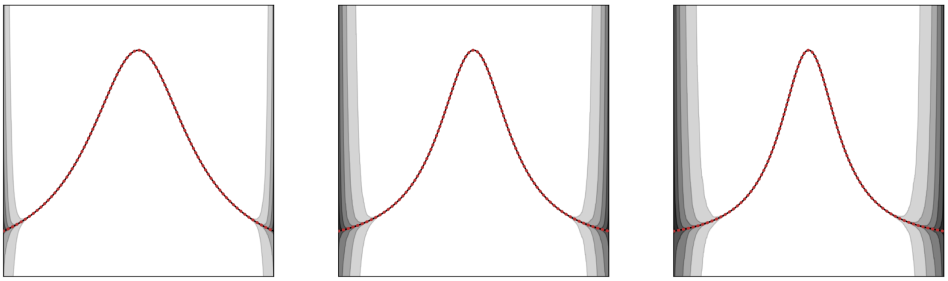

FIGURE 5. Probability distributions of the function values for empirical approximations of $f(x) = \frac{1}{1+cx^2}$ (red) by Legendre polynomials of degree 29. Different approximations correspond to different random draws of $n = 100$ sampling points from the uniform measure on $[-1, 1]$. (a) $c = 5$, mean error: 2×10^3. (b) $c = 10$, mean error: 2×10^5. (c) $c = 15$, mean error: 2×10^6.

assumptions on the model class as was done in [22]. We, however, believe that this problem can be handled by taking the regularity of the sought function u into account. Figure 5 provides first numerical evidence in support of this hypothesis. The model class used for all three experiments is the same and only the regularity of the function varies. Even though the best approximation error in all three cases is bounded by 10^{-3}, we can observe how the empirical approximations deteriorate with decreasing regularity. The relative errors for the empirical approximation increase from 10^{-2} to 10^{1}.

Despite the mentioned limitation, we obtain qualitatively similar results to those that are reported for more specialized approaches. In particular, this concerns the emergence of an optimal sampling measure in Section 3.1, the importance of weighted sparsity (rather than standard sparsity) in Section 3.2 and the advantage of multilevel sampling in Example 4.3. The generality of our theory also allows us to combine these result and derive optimal weight functions in the setting of weighted sparsity. Since these results rely only on an estimation of the probability of the RIP, they can be compared to results on weighted ℓ^1-minimization and we observe that using an optimal weight function can improve the quality of the estimate.

In a final section, the dependence of the sample complexity on the seminorm that is used is investigated. Expectedly, we observe faster convergence when stronger norms are used and provide a theoretical reason for this effect.

Despite several remaining problems, we believe that this work is a promising first step towards a general theory for the sample complexity of the nonlinear least squares problem. We also want to emphasise that, although our discussion is limited to well-known model classes, the developed theory can be applied to arbitrary model classes which may even be constructed empirically by methods such as manifold learning.

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

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

Using a union bound argument it follows that

$$\mathbb{P}\left[\max_{1\leq j\leq \nu} \left|\|u_j\|^2 - \|u_j\|_n^2\right| > \tfrac{\delta}{2}\right] \leq \sum_{1\leq j\leq \nu} \mathbb{P}\left[\left|\|u_j\|^2 - \|u_j\|_n^2\right| > \tfrac{\delta}{2}\right]$$

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

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

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

for each $1 \leq j \leq \nu$ by a standard concentration inequality. Combining both inequalities yields Lemma 2.7.

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 $\not{k}_y := \sqrt{\ell_y}$ and use the reverse triangle inequality

$$|\not{k}_y(u) - \not{k}_y(v)| \leq \not{k}_y(u - v) \leq \operatorname*{ess\,sup}_{y\in Y} \not{k}_y(u - v) = \|u - v\|_{w,\infty}.$$

Since $\not{k}_y$ is bounded by $\sqrt{K}$, we can use the Lipschitz continuity of $x \mapsto x^2$ on $[-\sqrt{K}, \sqrt{K}]$ to conclude

$$|\ell_y(u) - \ell_y(v)| \leq 2\sqrt{K}|\not{k}_y(u) - \not{k}_y(u)| \leq 2\sqrt{K}\|u - v\|_{w,\infty}.$$

□

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*

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

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

$$\Big|\|u\|^2 - \|u_j\|^2\Big| \leq \int_Y |\ell_y(u) - \ell_y(u_j)|w(y)^{-1}\,\mathrm{d}\rho(y) \leq 2\sqrt{K}\|u - u_j\|_{w,\infty} \leq \tfrac{\delta}{4}$$

and almost surely

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

Therefore, by triangle inequality,

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

Taking the maximum concludes the proof. □

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

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

Proof. By Lemma A.2

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

holds almost surely. In this event we know that

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

which concludes the proof. □

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*

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

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

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

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

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

We first need to show that $\hat{b}$ is measurable. For this let $\{u_j\}_{j=1}^\infty$ be a countable dense subset in A. Then

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

is the supremum over a countable set of measurable functions and as such measurable. The first assertion now follows by definition of K.

For the second assertion, we start by showing the integrability of $\hat{b}$ *via*

$$\int_Y \hat{b}(y)\,\mathrm{d}\rho(y) \leq \sup_{y\in Y} w(y) \sup_{v\in A}|v|_y^2 \int_Y w(y)^{-1}\,\mathrm{d}\rho(y) = \sup_{v\in A}\|v\|_{w,\infty}^2.$$

Now choose $R > 0$ such that $\|v\| \leq R$ for all $v \in A$. Since $\|v\|_{w,\infty} \leq \sqrt{K(A)}\|v\| \leq \sqrt{K(A)}R$ for all $v \in A$, we can conclude that $\hat{b}$ is integrable.

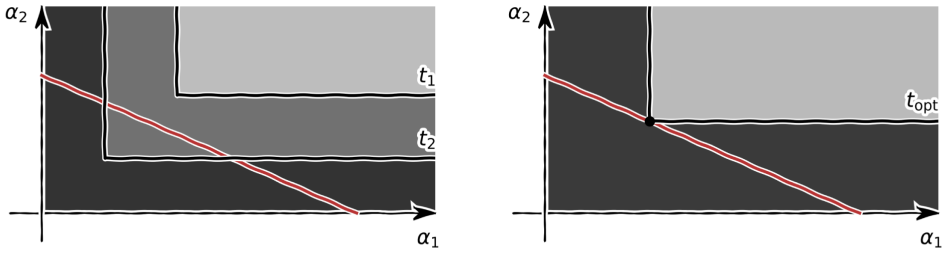

FIGURE B.1. The set of feasible α_1, α_2 satisfying $\alpha_1, \alpha_2 > 0$ and $\alpha_1 I_1 + \alpha_2 I_2 = 1$ is displayed in red. Contour lines $k(\alpha_1, \alpha_2) = t$ of the function $k(\alpha_1, \alpha_2) = \alpha_1^{-1} \vee \alpha_2^{-1}$ for $t_1 < t_2$ (*left*) and for the optimal value $t_{\text{opt}} = \alpha_1^{-1} = \alpha_2^{-1}$ (*right*) are drawn in black.

It remains to show that the weight function $w = \left\|\hat{b}\right\|_{L^1(Y,\rho)} \hat{b}^{-1}$ is indeed optimal. We only sketch the proof of this assertion. By substituting $w = \left(v\hat{b}\right)^{-1}$, the minimization problem

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

is equivalent to

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

which is a non-convex optimization problem under linear constraints. The second assertion is now equivalent to the statement that the minimal v is a constant function and the constraint $\int_Y \hat{b}v \,\mathrm{d}\rho = 1$ implies $w = \left\|\hat{b}\right\|_{L^1(Y,\rho)} \hat{b}^{-1}$.

To prove that a minimal v has to be constant, let $\Omega_1, \Omega_2 \subseteq Y$ be any disjoint subsets with positive measures and $\Omega_1 \cup \Omega_2 = Y$. Then v can be written as $v = \alpha_1 v_1 + \alpha_2 v_2$ with

$$\alpha_k := \left\|v^{-1}\right\|^{-1}_{L^\infty(\Omega_k,\rho)} \qquad \text{and} \qquad v_k := \frac{v\chi_{\Omega_k}}{\alpha_k} \qquad \text{for } k = 1, 2.$$

Now observe that

$$\left\|v^{-1}\right\|_{L^\infty(Y,\rho)} = \left\|v^{-1}\right\|_{L^\infty(\Omega_1,\rho)} \vee \left\|v^{-1}\right\|_{L^\infty(\Omega_2,\rho)} = \alpha_1^{-1} \vee \alpha_2^{-1}.$$

Moreover, $v > 0$ implies $\alpha_1, \alpha_2 > 0$ and the linear constraint can hence be written as $\alpha_1 I_1 + \alpha_2 I_2 = 1$ with $I_k := \int_Y \hat{b}v_k \,\mathrm{d}\rho$ for $k = 1, 2$. Since v is optimal, it must also satisfy

$$\min_{\alpha_1,\alpha_2} \alpha_1^{-1} \vee \alpha_2^{-1} \quad \text{s.t.} \quad \alpha_1, \alpha_2 > 0 \text{ and } \alpha_1 I_1 + \alpha_2 I_2 = 1.$$

Figure B.1 illustrates why the solution must be $\alpha_1 = \alpha_2$. This means that an optimal function v has to satisfy $\left\|v^{-1}\right\|_{L^\infty(\Omega_1,\rho)} = \left\|v^{-1}\right\|_{L^\infty(\Omega_2,\rho)}$. The claim now follows since the subsets Ω_1 and Ω_2 were chosen arbitrarily.

APPENDIX C. PROOF OF EQUATION (4.1)

Recall that $\mathscr{V} := H^m(Y, \rho)$ where $Y \subseteq \mathbb{R}^d$ is a Lipschitz domain and $A \subseteq \mathscr{H} := H^M(Y, \rho)$ and that the considered seminorm is given by

$$|v|_y^2 = \|L_y v\|_2^2 = \sum_{\substack{\alpha \in \mathbb{N}^d \\ |\alpha| \leq m}} \left|L_y^\alpha v\right|^2,$$

where the family of linear operators L_y^α is defined by $L_y^\alpha v = D^\alpha v(y)$ for all $y \in Y$ and all $\alpha \in \mathbb{N}^d$ with $|\alpha| \leq m$. Observe that $D^\alpha v \in H^\mu(Y,\rho)$ for all $\alpha \in \mathbb{N}^d$ with $|\alpha| \leq m$, where $\mu := M - m > \frac{d}{2}$. It was shown in [8] 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 $Y = \mathbb{R}^d$. In the following we compute

$$\kappa(y) = \|L_y\|^2_{\mathscr{L}\left(\mathscr{H},\mathbb{R}^{|\{\alpha\in\mathbb{N}^d:|\alpha|\leq m\}|}\right)} = \sum_{\substack{\alpha\in\mathbb{N}^d\\|\alpha|\leq m}} \left\|L_y^\alpha\right\|^2_{\mathscr{H}^*}.$$

As in [36] the Riesz representative of L_y^α,

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

can be obtained *via* the Fourier transform and some of its standard properties. Thus,

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

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

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

The multinomial theorem states that

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

As a consequence,

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

This leads to the estimate

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

The recurrence relation (2.147) in [23] together with $\mu > \frac{d}{2}$ yields

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

Consequently,

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

Acknowledgements. We thank the anonymous referees for suggestions that helped to significantly improve the manuscript and also to correct an error. We also thank Leon Sallandt, Mathias Oster and Michael Götte for fruitful discussions. 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).

References

[1] B. Adcock, Infinite-dimensional compressed sensing and function interpolation. *Found. Comput. Math.* **18** (2017) 661–701.
[2] B. Adcock and Y. Sui, Compressive hermite interpolation: sparse, high-dimensional approximation from gradient-augmented measurements. *Constr. Approx.* **50** (2019) 167–207.
[3] B. Adcock, A.C. Hansen, C. Poon and B. Roman, Breaking the coherence barrier: a new theory for compressed sensing. *Forum Math. Sigma* **5** (2017) e4.
[4] M. Bachmayr and R. Schneider, Iterative methods based on soft thresholding of hierarchical tensors. *Found. Comput. Math.* **17** (2017) 1037–1083.
[5] 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. *SIAM J. Math. Data Sci.* **2** (2020) 631–657.
[6] B. Bohn, On the convergence rate of sparse grid least squares regression. In: Sparse Grids and Applications-Miami 2016. Springer (2018) 19–41.
[7] M.F. Bugallo, V. Elvira, L. Martino, D. Luengo, J. Miguez and P.M. Djuric, Adaptive importance sampling: the past, the present, and the future. *IEEE Signal Process. Mag.* **34** (2017) 60–79.
[8] V. Burenkov, Extension theorems for sobolev spaces. In: The Maz'ya Anniversary Collection, edited by J. Rossmann, P. Takáč and G. Wildenhain. Birkhäuser Basel (1999).
[9] E.J. Candès and D.L. Donoho, New tight frames of curvelets and optimal representations of objects with piecewiseC2singularities. *Commun. Pure Appl. Math.* **57** (2003) 219–266.
[10] E.J. Candes and T. Tao, The power of convex relaxation: near-optimal matrix completion. *IEEE Trans. Inf. Theory* **56** (2010) 2053–2080.
[11] E.J. Candès, J.K. Romberg and T. Tao, Stable signal recovery from incomplete and inaccurate measurements. *Commun. Pure Appl. Math.* **59** (2006) 1207–1223.
[12] C. Chen, B. Zhang, A. Del Bue and V. Murino, Manifold constrained low-rank decomposition. In: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW) (2017) 1800–1808.
[13] 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: M2AN* **49** (2015) 815–837.
[14] A. Cohen and G. Migliorati, Optimal weighted least-squares methods. *SMAI J. Comput. Math.* **3** (2017) 181–203.
[15] F. Cucker and S. Smale, On the mathematical foundations of learning. *Bull. Am. Math. Soc.* **39** (2001) 1–50.
[16] F. Cucker and D.X. Zhou, Learning Theory: An Approximation Theory Viewpoint. *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press (2007).
[17] J. Daws, Jr., A. Petrosyan, H. Tran and C.G. Webster, A weighted ℓ_1-minimization approach for wavelet reconstruction of signals and images. Preprint arXiv:1909.07270 (2019).
[18] S. Dirksen, Tail bounds via generic chaining. *Electron. J. Probab.* **20** (2015) 1–29.
[19] K.-L. Du and M.N.S. Swamy, Compressed Sensing and Dictionary Learning. Springer London, London (2019) 525–547.
[20] M. Eigel, R. Schneider, P. Trunschke and S. Wolf, Variational Monte Carlo – bridging concepts of machine learning and high-dimensional partial differential equations. *Adv. Comput. Math.* **45** (2019) 2503–2532.
[21] Y.C. Eldar and G. Kutyniok, Compressed Sensing: Theory and Applications. Cambridge University Press (2012).
[22] A. Goeßmann, M. Götte, I. Roth, R. Sweke, G. Kutyniok and J. Eisert, Tensor network approaches for learning non-linear dynamical laws. Preprint arXiv:2002.12388 (2020).
[23] I.S. Gradshteyn, I.M. Ryzhik and D.F. Hays, Table of Integrals, Series, and Products. Academic Press (2014).

[24] L. Grasedyck and W. Hackbusch, An introduction to hierarchical (H-) rank and TT-rank of tensors with examples. *Comput. Methods Appl. Math.* **11** (2011) 291–304.
[25] L. Grasedyck and S. Krämer, Stable ALS approximation in the TT-format for rank-adaptive tensor completion. *Numer. Math.* **143** (2019) 855–904.
[26] L. Györfi, M. Kohler, A. Krzyżak and H. Walk, A Distribution-Free Theory of Nonparametric Regression. Springer, New York (2002).
[27] W. Hackbusch, Tensor Spaces and Numerical Tensor Calculus. Vol. 42. Springer Science & Business Media (2012).
[28] F.L. Hitchcock, The expression of a tensor or a polyadic as a sum of products. *J. Math. Phys.* **6** (1927) 164–189.
[29] A. Jung, Y.C. Eldar and N. Görtz, On the minimax risk of dictionary learning. *IEEE Trans. Inf. Theory* **62** (2016) 1501–1515.
[30] E. Kowalski, Pointwise bounds for orthonormal basis elements in hilbert spaces (2011).
[31] G. Kutyniok, P. Petersen, M. Raslan and R. Schneider, A theoretical analysis of deep neural networks and parametric PDEs. *Constr. Approx.* (2021) DOI: 10.1007/s00365-021-09551-4.
[32] Q. Meng, X. Xiu and Y. Li, Manifold constrained low-rank and joint sparse learning for dynamic cardiac MRI. *IEEE Access* **8** (2020) 142622–142631.
[33] G. Migliorati, F. Nobile, E. von Schwerin and R. Tempone, Analysis of discrete l^2 projection on polynomial spaces with random evaluations. *Found. Comput. Math.* **14** (2014) 419–456.
[34] G. Migliorati, F. Nobile and R. Tempone, Convergence estimates in probability and in expectation for discrete least squares with noisy evaluations at random points. *J. Multivariate Anal.* **142** (2015) 167–182.
[35] NIST Digital Library of Mathematical Functions.
[36] 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. *Anal. App.* **16** (2018) 693–715.
[37] P. Petersen, Shearlet approximation of functions with discontinuous derivatives. *J. Approx. Theory* **207** (2016) 127–138.
[38] H. Rauhut, Compressive sensing and structured random matrices. *Theor. Found Numer. Methods Sparse Recover.* **9** (2010) 1–92.
[39] H. Rauhut and R. Ward, Interpolation via weighted $\ell 1$ minimization. *Appl. Comput. Harmonic Anal.* **40** (2016) 321–351.
[40] H. Rauhut, R. Schneider and Ž. Stojanac, Low rank tensor recovery via iterative hard thresholding. *Linear Algebra App.* **523** (2017) 220–262.
[41] Y. Traonmilin and R. Gribonval, Stable recovery of low-dimensional cones in hilbert spaces: one RIP to rule them all. *Appl. Comput. Harmonic Anal.* **45** (2018) 170–205.
[42] J.A. Tropp, User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.* **12** (2012) 389–434.
[43] V.N. Vapnik and A.Y. Chervonenkis, Necessary and sufficient conditions for the uniform convergence of means to their expectations. *Theory Prob. App.* **26** (1982) 532–553.
[44] 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). IEEE (2009) 189–192.
[45] M. Yuan and C.-H. Zhang, On tensor completion via nuclear norm minimization. *Found. Comput. Math.* **16** (2015) 1031–1068.