*SMAI-JCM*
SMAI Journal of Computational Mathematics

# Weighted sparsity and sparse tensor networks for least squares approximation

Philipp Trunschke, Martin Eigel & Anthony Nouy

Volume 11 (2025), p. 289-333.
https://doi.org/10.5802/smai-jcm.126

© The authors, 2025.

*The SMAI Journal of Computational Mathematics is a member of the Centre Mersenne for Open Scientific Publishing*
http://www.centre-mersenne.org/
Submissions at https://smai-jcm.centre-mersenne.org/ojs/submission
e-ISSN: 2426-8399

*SMAI Journal of Computational Mathematics*
Vol. 11, 289-333 (2025)

# Weighted sparsity and sparse tensor networks for least squares approximation

Philipp Trunschke [1]
Martin Eigel [2]
Anthony Nouy [3]

[1] Centrale Nantes, Nantes Université, Laboratoire de Mathématiques Jean Leray, CNRS UMR 6629, France
*E-mail address*: philipp.trunschke@univ-nantes.fr
[2] Weierstrass Institute, Mohrenstrasse 39, 10117 Berlin, Germany
*E-mail address*: martin.eigel@wias-berlin.de
[3] Centrale Nantes, Nantes Université, Laboratoire de Mathématiques Jean Leray, CNRS UMR 6629, France
*E-mail address*: anthony.nouy@ec-nantes.fr.

**Abstract.** Approximation of high-dimensional functions is a problem in many scientific fields that is only feasible if advantageous structural properties, such as sparsity in a given basis, can be exploited. A relevant tool for analysing sparse approximations is Stechkin's lemma.

In its standard form, however, this lemma does not allow to explain convergence rates for a wide range of relevant function classes. This work presents a new weighted version of Stechkin's lemma that improves the best $n$-term rates for weighted $\ell^p$-spaces and associated function classes such as Sobolev or Besov spaces. For the class of holomorphic functions, which occur as solutions of common high-dimensional parameter-dependent PDEs, we recover exponential rates that are not directly obtainable with Stechkin's lemma.

Since weighted $\ell^p$-summability induces weighted sparsity, compressed sensing algorithms can be used to approximate the associated functions. To break the curse of dimensionality, from which these algorithms suffer, we utilise that sparse approximations can be encoded efficiently using tensor networks with sparse component tensors. We also demonstrate that weighted $\ell^p$-summability induces low ranks, which motivates a second tensor train format with low ranks and a single weighted sparse core. We present new alternating algorithms for best $n$-term approximation in both formats.

To analyse the sample complexity for the new model classes, we derive a novel result of independent interest that allows the transfer of the restricted isometry property from one set to another sufficiently close set. Although they lead up to the analysis of our final model class, our contributions on weighted Stechkin and the restricted isometry property are of independent interest and can be read independently.

**2020 Mathematics Subject Classification.** 15A69, 41A30, 62J02, 65Y20, 68Q25.

**Keywords.** least squares, sample efficiency, sparse tensor networks, alternating least squares.

## 1. Introduction

Approximating an unknown function from data is a fundamental problem in computational science and machine learning. In many applications, the sought function may depend on a large number of parameters, rendering the approximation task susceptible to the *curse of dimensionality* (CoD), i.e. an exponential complexity in the dimension of the problem or the amount of sample points required to obtain an accurate approximation. This is particularly problematic when the amount of data available is limited due to practical constraints. Nevertheless, many practically relevant functions can be

This project is funded by the ANR-DFG project COFNET (ANR-21-CE46-0015). ME acknowledges the partial support by the DFG SPP 2298 (Mathematics of Deep Learning). This work was partially conducted within the France 2030 framework programme, Centre Henri Lebesgue ANR-11-LABX-0020-01.
https://doi.org/10.5802/smai-jcm.126
© The authors, 2025

approximated efficiently using a judiciously chosen set of functions. Given a function that can be well approximated by a sparse expansion on some basis, results in compressive sensing guarantee an accurate approximation from a small number of sample points. The required sets (or dictionaries) can be found by exploiting regularity properties of the sought function. A common characterisation of smooth functions is given in terms of the decay of their Fourier series. This can be viewed as promoting a structured sparsity where low-order Fourier modes are more likely to contribute to the total norm of the function. As a consequence, smooth functions admit approximately sparse representations, enabling an efficient numerical reconstruction.

Another form of low-dimensional structure induced by smoothness is low-rank approximability. This structure is exploited e.g. in reduced basis methods or proper orthogonal decomposition [17, 41] and in a more general form in hierarchical tensor formats such as the popular tensor trains (TT) [43].

The aim of this paper is to develop a sparse approximation algorithm which simultaneously exploits sparsity and low-rank properties, thus enabling an efficient approximation of large function sets. Central tools for this are a new weighted version of the well-known Stechkin's lemma and a novel analysis of the restricted isometry property which allows to accommodate any space which can be approximated by weighted sparse expansions. These developments should be of independent interest in the study of sparse and general nonlinear least squares approximations. We carry out the theoretical analysis of convergence rates for weighted $\ell^p$ functions. Moreover, we present representations of these sparse vectors (or sequences) in sparse low-rank formats and discuss the application to parametric PDEs.

This is not the first work that proposes the utilisation of sparsity in the component tensors of a tensor network. In [26] and [40], the authors consider the abstract setting of empirical risk minimisation on bounded model classes of potentially sparse tensor networks. They present model selection strategies for the network topology and sparsity pattern. Due to the use of empirical risk minimisation, they obtain standard error bounds for arbitrary risk functions satisfying boundedness and Lipschitz continuity assumptions (cf. [40], which relies on approximation results from [3, 4]). However, in the case of least squares risk, these strong assumptions restrict the application to bounded model classes. Moreover, they do not guarantee an equivalence of errors, which translates to the error decreasing with a slow Monte Carlo rate. This is intolerable when striving for small relative errors, which is often the case in numerical schemes.

In [16] the authors propose an algorithm that computes a sparse best approximation in the model class of sparse rank-1 tensors. Conceptually, this algorithm is very similar to our Algorithm 2 but is restricted to a sum of unweighted sparse rank-1 tensors. The restriction to a sum of rank-1 tensors implies a suboptimal convergence with respect to the rank and the use of unweighted sparsity means that, in the worst case, vastly more sample points may be required than are actually necessary.

A very similar ($s^2$-PGD) method is also proposed in [47]. The method optimises with regard to the same model class as our Algorithm 2 but does not orthogonalise the component tensors in between the micro steps. It is not clear if this lack of orthogonalisation can result in numerical instabilities as it would in the classical ALS method. Moreover, the lack of orthogonalisation prevents the use of the correct weight sequence in the micro steps of their sparse ALS and does not allow for the same automatic rank adaptation as our algorithm.

Finally, block-sparse tensor networks are a well-known tool in the numerics of quantum mechanics [51] and were recently introduced to the mathematics community by [8]. This theory is already used in [24] to perform least squares regression in a model class of tensor trains restricted to subspaces of homogeneous polynomials of fixed degree. The basis selection performed in our second algorithm is conceptually very similar to the restriction to eigenspaces used in [8] to ensure block-sparsity. We believe that our algorithm can be interpreted as a generalisation of the regression on block-sparse tensor trains. In contrast to this approach, where the sparsity structure has to be known in advance, our algorithms explores the sparsity automatically.

### 1.1. **Weighted sparsity**

Sparse approximability of a function can be expressed by the $\ell^q$-summability of the coefficients sequence of its basis or frame representation. The following central result, commonly attributed to Stechkin [18, 21], is a key tool to provide convergence rates for sparse approximations.

**Lemma 1.1** (Stechkin)**.** *Let $0 < q < p \leq \infty$ and let $\boldsymbol{v} \in \ell^q$. Define $J_n$ as the set of indices corresponding to the $n$ largest elements of the sequence $|\boldsymbol{v}|$ and $P_{J_n}\boldsymbol{v} = \sum_{j\in J_n} \boldsymbol{v}_j \boldsymbol{e}_j$, where $\boldsymbol{e}_j$ is the sequence with 1 at index $j$ and 0 everywhere else. Then*

$$\|\boldsymbol{v} - P_{J_n}\boldsymbol{v}\|_{\ell^p} \leq (n+1)^{-s}\|\boldsymbol{v}\|_{\ell^q}, \quad s := \frac{1}{q} - \frac{1}{p}.$$

Given a separable Banach space $V$ of functions with a normalized basis (or frame) $\{B_k\}_{k\in\mathbb{N}} \subseteq V$, each element $u \in V$ can be identified with its coefficient sequence $\boldsymbol{u}$ with respect to this basis. Lemma 1.1 hence yields the convergence estimate for the best $n$-term approximation $u_n$ of $u$:

$$\|u - u_n\|_V \leq \|\boldsymbol{u} - P_{J_n}\boldsymbol{u}\|_{\ell^p} \leq (n+1)^{-s}\|\boldsymbol{u}\|_{\ell^q}, \quad s = 1/q - 1/p,$$

for $p = 1 > q$ in the general Banach case, or $p = 2 > q$ when $V$ is a Hilbert space and $\{B_k\}_{k\in\mathbb{N}}$ is an orthonormal basis of $V$. A disadvantage of the standard Stechkin's lemma is that it can only predict algebraic approximation rates, and these rates are suboptimal for some relevant classes of functions.

**Contributions.** To overcome these issues, we introduce for any sequence $\boldsymbol{\omega} \in [0,\infty]^{\mathbb{N}}$ the $\boldsymbol{\omega}$-weighted sequence space

$$\ell^q_{\boldsymbol{\omega}} := \{\boldsymbol{v} \in \mathbb{R}^{\mathbb{N}} : \|\boldsymbol{v}\|_{\ell^q_{\boldsymbol{\omega}}} := \|\boldsymbol{\omega}\boldsymbol{v}\|_{\ell^q} < \infty\},$$

where $\boldsymbol{\omega}\boldsymbol{v} = (\boldsymbol{\omega}_\nu \boldsymbol{v}_\nu)_\nu$ denotes the element-wise multiplication of the two sequences $\boldsymbol{\omega}$ and $\boldsymbol{v}$. With these spaces, a corresponding weighted version of Stechkin's lemma is derived, which enables to better exploit classical regularity in terms of convergence rates, significantly improving results of the classical lemma. Indeed, in Section 2 we recall that the weighted $\ell^q_{\boldsymbol{\omega}}$-spaces correspond to a variety of function spaces such as Barron, Besov and Sobolev spaces. This makes possible to relate the summability $\boldsymbol{u} \in \ell^q$ directly to more natural regularity assumptions such as $u$ being in the Sobolev space $H^k$ (cf. Example 2.12). We compare the obtained results to previous works and discuss the relation of the weighted $\ell^p$ spaces to unweighted and monotone $\ell^p$ spaces (cf. [1]). If the employed weight sequence increases super-algebraically, the new weighted bound has a significantly faster decay than the corresponding unweighted bound. Moreover, there are also improvements for algebraically increasing weight sequences, namely a significant reduction of multiplicative constants in the approximation estimate. Because of this, the analysis provides immediate convergence bounds even in the non-asymptotic setting, i.e. for finite-dimensional linear spaces.

### 1.2. **Application to parametric partial differential equations**

The numerical solution of high-dimensional parametric operator equations has become a highly active research field in the last decade, particularly in the area of Uncertainty Quantification (UQ) and in relation to modern (scientific) machine learning, see for instance [17, 50] and references therein. The parameter domain is often high- or even infinite-dimensional, making it computationally challenging to approximate the solution in linear spaces due to the CoD. It hence is mandatory to exploit structural properties of the respective functions. When relying on sparsity as we do, the required summability constraints can be deduced from smoothness. Indeed, it is shown in Section 2 that weighted summability with algebraically increasing weight sequences can often be derived from standard regularity assumptions. Morevover, certain assumptions on the data allow us to derive even stronger summability properties for the solution of parametric PDEs as shown e.g. in [5, 6].

We recall a prototypical parametric linear second order elliptic problem and its solution properties as a motivation for the results of this work. For a given bounded Lipschitz domain $D \subseteq \mathbb{R}^d$ with $d \in \mathbb{N}$ and some source function $f \in L^2(D)$, consider the linear elliptic PDE

$$\begin{aligned} -\operatorname{div}_x(a(x,y)\nabla_x u(x,y)) &= f(x), \quad \text{in } D, \\ u(x,y) &= 0, \qquad \text{on } \partial D, \end{aligned} \tag{1.1}$$

where $y \in \mathbb{R}^L$ is a high-dimensional ($L \in \mathbb{N}$) or infinite-dimensional ($L = \infty$) parameter vector determining the coefficient field $a$ and hence the solution $u$. With typical applications in modelling stochastic flow through porous media (such as groundwater flow [56]), the diffusion coefficient is often defined by a Karhunen–Loève type expansion [39, 52], which can be constructed to represent random fields with bounded variance and typically takes the form

$$a(x,y) = \sum_{j=1}^{L} a_j(x) y_j + a_0(x) \qquad \text{with } y \sim \mathcal{U}([-1,1])^{\otimes L} \qquad \text{or} \tag{1.2}$$

$$\ln(a(x,y)) = \sum_{j=1}^{L} a_j(x) y_j \qquad \text{with } y \sim \mathcal{N}(0,1)^{\otimes L}. \tag{1.3}$$

In these applications, the functions $a_j : D \to \mathbb{R}$ are scaled $L^2(D)$-orthogonal eigenfunctions of the covariance operator of $a$ or $\ln(a)$. This specific choice is not necessary for the application of our theory and other more advantageous expansions (cf. [7]) may be considered as well.

We recall some results from [5, 6] on the analysis and approximation of the parameter-to-solution map

$$y \mapsto u(y) := u(\bullet, y),$$

induced by the model (1.1)–(1.3). For (1.2), the following result was shown recently.

**Theorem 1.2** (Theorem 3.1 in [6])**.** *Consider problem* (1.1) *with affine coefficients* (1.2)*. Assume that there exists a sequence* $\boldsymbol{\rho} \in (1,\infty)^L$ *such that*

$$\sup_{x \in D} \frac{1}{a_0} \sum_{j \in [L]} \boldsymbol{\rho}_j |a_j(x)| < 1.$$

*Then the map* $y \mapsto u(y)$ *belongs to* $L^k([-1,1]^L, \mathrm{d}\gamma; H_0^1(D))$ *for all* $k \in \mathbb{N} \cup \{\infty\}$*, where* $\gamma$ *is the uniform measure. Hence, there exists an expansion of* $u(x,y) = \sum_{\nu \in \mathcal{F}} u_\nu(x) L_\nu(y)$ *in terms of Legendre polynomials* $(L_\nu)_{\nu \in \mathcal{F}}$*, where* $\mathcal{F}$ *is the set of multi-indices in* $\mathbb{N}^L$ *with finite support. Moreover, the sequence of coefficients satisfies*

$$\sum_{\nu \in \mathcal{F}} \boldsymbol{\omega}_\nu^2 \|u_\nu\|_{H_0^1(D)}^2 < \infty,$$

*with* $\boldsymbol{\omega}_\nu := \prod_{j \in [L]} (2\nu_j + 1)^{-1/2} \boldsymbol{\rho}_j^{\nu_j}$.

For the case of unbounded Gaussian parameters (1.3), we recall the following result.

**Theorem 1.3** (Theorems 2.2, 3.3 and 4.2 in [5])**.** *Consider the model* (1.1) *with affine coefficients* (1.3)*. Assume that there exists an* $r \in \mathbb{N}$ *and a sequence* $\boldsymbol{\rho} \in (0,\infty)^L$ *such that*

$$\sup_{x \in D} \sum_{j \in [L]} \boldsymbol{\rho}_j |a_j(x)| < \frac{\ln 2}{\sqrt{r}} \qquad \textit{and} \qquad \sum_{j \in [L]} \exp(-\boldsymbol{\rho}_j^2) < \infty.$$

*Then the map* $y \mapsto u(y)$ *belongs to* $L^k(\mathbb{R}^L, \mathrm{d}\gamma; H_0^1(D))$ *for all* $k \in \mathbb{N}$*, where* $\gamma$ *is the Gaussian measure on* $\mathbb{R}^L$*. Hence, there exists an expansion of* $u(x,y) = \sum_{\nu \in \mathcal{F}} u_\nu(x) H_\nu(y)$ *in terms of Hermite polynomials* $(H_\nu)_{\nu \in \mathcal{F}}$*, where* $\mathcal{F}$ *is the set of multi-indices in* $\mathbb{N}^L$ *with finite support. Moreover, the sequence*

*of coefficients satisfies*

$$\sum_{\nu\in\mathcal{F}} \omega_\nu^2 \|u_\nu\|_{H_0^1(D)}^2 < \infty,$$

*with* $\boldsymbol{\omega}_\nu := \prod_{j\in[L]}(\sum_{l=0}^r \binom{\nu_j}{l}\boldsymbol{\rho}_j^{2l})^{1/2}$.

**Contributions.** Using the weighted sequence spaces and Stechkin's lemma, we propose an alternative method of proof for the summability of the solution of parametric PDEs in Section 3. We show that in the case of a single parameter ($L = 1$), these summability properties already follow from the analyticity of the parameter to solution map. The new derivation only relies on the weighted version of Stechkin's lemma and elementary techniques. This results in similar bounds to those in Theorem 1.2 and 1.3 with exponential decay of the basis coefficients.

### 1.3. Numerical methods for weighted sparse approximation using sparse tensor train format

Equally important as the approximation error analysis is the availability of actual computational methods. For a probability measure $\gamma$ on some set $Y$, let $\mathcal{M} \subseteq L^2(Y,\gamma)$ be a *model class* of functions in which $u \in \mathcal{V}$ should be approximated. Defining the norms

$$\|\bullet\| := \|\bullet\|_{L^2(Y,\gamma)} \qquad \text{and} \qquad \|\bullet\|_{w,\infty} := \|w^{1/2}\bullet\|_{L^\infty(Y,\gamma)}, \tag{1.4}$$

where $w$ is some positive *weight function*, the problem of determining the best-approximation of $u$ in $\mathcal{M}$ can be formulated as

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

Since the $L^2$-norm cannot be computed exactly in high-dimensional settings, a popular remedy is to introduce an empirical estimator with samples $\boldsymbol{y} := \{y^i\}_{i=1}^n$ and the respective weighted least-squares minimisation, namely

$$u_{\mathcal{M},n} \in \operatorname*{arg\,min}_{v\in\mathcal{M}} \|u - v\|_n \quad \text{with} \quad \|v\|_n := \left(\frac{1}{n}\sum_{i=1}^n w(y^i)|v(y^i)|^2\right)^{1/2}. \tag{1.5}$$

A natural choice is to take $w$ such that $\int_Y w^{-1}\,\mathrm{d}\gamma = 1$, and to draw the points $y_i$ independently from the measure $w^{-1}\gamma$ for all $i = 1,\dots,n$. This approach results in approximations with guaranteed error bounds when assuming the *restricted isometry property* (RIP) that is known from compressed sensing [2, 12]. It is defined for a given set of functions $A$ by

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

If satisfied for a parameter $\delta \in (0,1)$, the error of estimator (1.5) can be bounded as follows.

**Proposition 1.4** (Theorem 3 in [54]). *If* $\mathrm{RIP}_{\{u_{\mathcal{M}}\}-\mathcal{M}}(\delta)$ *holds, then*

$$\|u - u_{\mathcal{M},n}\| \le \|u - u_{\mathcal{M}}\| + \tfrac{2}{\sqrt{1-\delta}}\|u - u_{\mathcal{M}}\|_{w,\infty}\ .$$

The assumption of $\mathrm{RIP}_{\{u_{\mathcal{M}}\}-\mathcal{M}}(\delta)$ is a weaker version of the standard assumption $\mathrm{RIP}_{\mathcal{M}-\mathcal{M}}(\delta)$, which is often used when considering nested sequences $(\mathcal{M}_r)_{r\ge1}$ of model classes, like $r$-sparse vectors or rank-$r$ tensors, which satisfy $\mathcal{M}_r - \mathcal{M}_r \subseteq \mathcal{M}_{2r}$. We show that such a nestedness property is also satisfied for the model classes of sparse low-rank tensors considered in this paper.

Because $\mathrm{RIP}_{\mathcal{M}}(\delta)$ is a random event, a sufficient number of sample points has to be used to guarantee that it holds true. For theoretical reasons and since obtaining new sample points may be costly, a practical goal is to achieve this property with a minimal number $n$.

To leverage the existing results from least squares methods [19] in the development of numerical methods, one may rely on explicit bounds on the coefficients or a weighted summability property of the form $\boldsymbol{u} \in \ell^2_{\boldsymbol{\omega}}$. Given such a bound, we may define the sets $\Lambda_n$ corresponding to the $n$ smallest weights and prove that

$$\|(I - P_{\Lambda_n})\boldsymbol{u}\| \lesssim n^{-s},$$

where $s$ relates to the summability of the sequence $\boldsymbol{\omega}$. An example of this can be seen in [20]. From a statistical point of view, this has the advantage that there exist bounds that guarantee that a small number of parameter evaluations is sufficient to result in a quasi-best sparse approximation with high probability. Finding sets $\Lambda_n$ with the prescribed error bounds, however, relies on the knowledge of an exponentially increasing weight sequences $\boldsymbol{\omega}$. Such sequences do not exist for every PDE and their existence is not always easy to prove. Moreover, due to the reliance on optimal sampling, the cited work also requires the ability to draw new samples from a problem adapted measure.

An alternative approach that mitigates these issues is the use of weighted sparsity [11, 45]. Let $\boldsymbol{v}$ denote the sequence of coefficients of $v \in \mathcal{V}$ with respect to a given basis. Then the set of $\boldsymbol{\omega}$-weighted $r$-sparse sequences is given by the ball

$$B_{\ell^0_{\boldsymbol{\omega}}}(0,r) = \{\boldsymbol{v} \,:\, \|\boldsymbol{v}\|_{\ell^0_{\boldsymbol{\omega}}} \leq r\}, \tag{1.6}$$

where the $\ell^0_{\boldsymbol{\omega}}$-"norm" is a generalisation of the standard $\ell^0$-"norm" and is defined in Section 2. In [46] the authors show that a significantly improved bound for the probability of the RIP of $B_{\ell^0_{\boldsymbol{\omega}}}(0,r)$ can be derived when the $\ell^0$-"norm" is replaced by its weighted version. Although the shown a priori convergence rates still rely on weighted summability assumptions, the method itself does not. The only requirement is an upper bound on the $L^\infty$-norm of the basis functions. As a consequence, it can be applied easily in practice and is less reliant on the rate of decay of the sequence $\boldsymbol{\omega}$. The following theorem is a slight generalisation of this result.

**Theorem 1.5.** *Fix parameters $\delta, p \in (0,1)$. Let $\{B_j\}_{j\in[D]}$ be orthonormal in $L^2(Y,\gamma)$ and let $w \geq 0$ be any weight function satisfying $\|w^{-1}\|_{L^1(Y,\gamma)} = 1$. Assume the weight sequence $\boldsymbol{\omega}$ in the definition* (1.6) *of the model class $\mathcal{M}$ satisfies $\boldsymbol{\omega}_j \geq \|w^{1/2}B_j\|_{L^\infty(Y,\gamma)}$ and fix*

$$n \geq C\delta^{-2} r \max\{\log^3(r)\log(D), -\log(p)\}.$$

*Let $y_1, \dots, y_n$ be drawn independently from $w^{-1}\gamma$. Then the probability of* $\mathrm{RIP}_{B_{\ell^0_{\boldsymbol{\omega}}}(0,r)}(\delta)$ *exceeds $1-p$.*

**Proof.** To make the weight function $w$ that is used explicit, we define $\|v\|_{n,w} := \frac{1}{n}\sum_{i=1}^n w(y_i)v(y_i)^2$. Applying Theorem 5.2 from [46] to the $L^2(Y, w^{-1}\gamma)$-orthonormal basis $\{\psi_j := w^{1/2}B_j\}_{j\in[D]}$ shows that the probability of the event

$$\forall\, v \in \ell^2 : \|v\|_{\ell^0_{\boldsymbol{\omega}}} \leq s$$

exceeds $1-p$, which implies that

$$(1-\delta)\|\psi^\intercal v\|^2_{L^2(Y,w^{-1}\gamma)} \leq \|\psi^\intercal v\|^2_{n,1} \leq (1+\delta)\|\psi^\intercal v\|^2_{L^2(Y,w^{-1}\gamma)}$$

holds with probability higher than $1-p$. The claim follows, since $\|\psi^\intercal v\|_{L^2(Y,w^{-1}\gamma)} = \|B^\intercal v\|_{L^2(Y,\gamma)}$ and $\|\psi^\intercal v\|_{n,1} = \|B^\intercal v\|_{n,w}$. ∎

The index set $J_n$ that contains the $n$ largest coefficients of the function can in principle contain arbitrary large indices. This is an issue for algorithmic realisations, where $J_n$ must be restricted to lie in a finite set of candidate indices $\Lambda$. The set $\Lambda$ should be large enough to ensure that $J_n \subseteq \Lambda$ but not too large as to blow up the time complexity of the numerical algorithm, which scales at least linearly with $|\Lambda|$. Specifically, it has to be chosen carefully not to re-introduce the CoD. Without assumptions on the summability of the coefficients, such a set is difficult to find. For the model problems considered in this work, this is not a problem since appropriate candidate sets $\Lambda$ can be designed based on the

summability conditions in Theorem 1.2. For other problems such conditions are not known, which impedes the application of compressed sensing algorithms.

**Contributions.** Without prior knowledge, the exponentially large candidate set $\Lambda = \{0, \dots, d-1\}^M$ is a natural choice but classical algorithms for sparse approximation would yield a complexity polynomial in $|\Lambda| = d^M$, hence the CoD. We propose to use tensor trains [33, 43] to alleviate this CoD. Building upon results from [38], we show that the best $n$-term approximation can be represented in a sparse tensor train format with rank $n$. A sparse version of the ALS algorithm, which we call SALS, can be used to optimise over the sparse components. The complexity becomes polynomial in $n$ and $d$ and linear in $M$. This allows almost the same sample complexity bounds as in Theorem 1.5 while also admitting an admissible algorithmic realisation. We emphasise that this new algorithm is a feasible alternative to sparse approximation algorithm, not for the approximation in low-rank tensor format.

### 1.4. **Numerical methods for weighted sparse and low-rank tensor train approximation**

The results of Section 4 provide an approach to express sparse tensors as TTs with a rank that is bounded by the number of nonzero entries of the component tensors. We demonstrate that tensors with weighted sparsity are not only sparse but have also low rank, which is not exploited by the model class of sparse tensor trains from the previous section. Due to the special structure of these sparse tensor trains, the basis of every core tensor is strongly overparameterised. As a result, the linear systems arising in the microsteps of the SALS may become very large and the optimisation becomes very costly. Another consequence of this overparameterisation is that the sample size that is required for an accurate microstep is larger than it would have to be if only a minimal basis would have been used.

**Contributions.** As a possible solution to the overparametrisation issue we propose to round the sparse tensor back to minimal rank. Although this destroys the sparsity of the orthogonal component tensors, it retains the weighted sparsity of the core tensor. This yields a new model class of sparse and low-rank tensor trains. Investigating the probability of the RIP for this hybrid model class is the focus of Section 5. To do this, we show in Theorem 5.9 that the RIP on $B_{\ell^0_{\boldsymbol{\omega}}}(0, r)$ induces a RIP for any model class that is close enough with respect to an appropriate distance. This is a promising novel result that applies to any model class. In particular, we show in Theorem 5.12 that our hybrid model class satisfies the conditions of Theorem 5.9. It thereby inherits the RIP from sparse vectors that is guaranteed by the stability result in Theorem 1.5. The results improve upon the previously developed theory for tensor reconstruction of solutions of high-dimensional parametric PDEs as presented in [53, 54].

## 2. **A weighted version of Stechkin's lemma**

In what follows, we are concerned with coefficient sequences $\boldsymbol{x}$ indexed by a set $\Lambda$, which may be finite or countably infinite. If not specified otherwise, we always assume that $\Lambda = \mathbb{N}$. Since any operation defined on the coefficients can be extended to an element-wise operation defined on sequences, we write e.g. $(\boldsymbol{xy})_j = \boldsymbol{x}_j \boldsymbol{y}_j$ as the element-wise product of two sequences $\boldsymbol{x}$ and $\boldsymbol{y}$ and $(\boldsymbol{x}/\boldsymbol{y})_j = \boldsymbol{x}_j / \boldsymbol{y}_j$ as the element-wise division. For any such sequence $\boldsymbol{x}$ and any subset $J \subseteq \Lambda$, define $P_J \boldsymbol{x}$ via

$$(P_J \boldsymbol{x})_j := \begin{cases} \boldsymbol{x}_j & j \in J \\ 0 & j \in \Lambda \setminus J. \end{cases}$$

In other words, $P_J$ is the canonical projection onto the linear space $\operatorname{span}\{\boldsymbol{e}_j : j \in J\}$, with $\boldsymbol{e}_j$ the canonical sequence having 1 at index $j$ and 0 everywhere else. Moreover, let $\operatorname{supp}(\boldsymbol{x}) := \{j \in \Lambda : \boldsymbol{x}_j \neq 0\}$ denote the support of $\boldsymbol{x}$. To a vector $\boldsymbol{\omega} \in [0,\infty]^\Lambda$ of weights we associate for $0 < p \leq \infty$ the weighted $\ell^p$ spaces

$$\ell^p_{\boldsymbol{\omega}} := \Big\{\boldsymbol{x} \in \mathbb{R}^\Lambda \ : \ \|\boldsymbol{x}\|_{\ell^p_{\boldsymbol{\omega}}} := \|\boldsymbol{\omega}\boldsymbol{x}\|_{\ell^p} < \infty\Big\}.$$

Central to our analysis is the weighted $\ell_0$-"norm" given by

$$\|\boldsymbol{x}\|_{\ell^0_{\boldsymbol{\omega}}} = \sum_{j\in\operatorname{supp}(\boldsymbol{x})} \boldsymbol{\omega}_j^2,$$

which counts the squared weights of the non-zero entries of $\boldsymbol{x}$. When $\boldsymbol{\omega} \equiv 1$, these weighted norms reproduce the standard $\ell^p$ norms.

In sparse approximation theory, the assumption $\boldsymbol{v} \in \ell^q$ is often central for the analysis. However, it provides no guarantee for the position of the largest elements in the sequence. For the purpose of numerical discretisation this is problematic since a truncation after the first $n$ terms of the sequence is not guaranteed to contain the largest elements. Without an explicit bound on the decay of $\boldsymbol{v}$, no bounds for the discretisation error can be given. We hence argue that it is natural to require an ordering of the terms that is induced by a weight sequence $\boldsymbol{\omega}$. Such a weighting exists for instance in the coefficients of solutions of parametric PDEs [5, 6]. Moreover, we will show later that such a weighting occurs naturally for many classical regularity classes like Sobolev and Besov spaces and for certain bases.

A very elegant proof of Stechkin's Lemma (Lemma 1.1) is provided in [17, Lemma 3.6], which relies on a basic bound for the decay of any $\boldsymbol{v} \in \ell^q$ and an application of Hölder's inequality. The same reasoning can be applied to obtain a proof for the Stechkin inequality in the weighted setting below.

**Lemma 2.1.** *Let $0 < q < p \leq \infty$ and $\boldsymbol{\alpha}, \boldsymbol{\sigma}, \boldsymbol{\omega} \in [0,\infty]^{\mathbb{N}}$ be sequences satisfying $\boldsymbol{\alpha}^p = \boldsymbol{\sigma}^{p-q}\boldsymbol{\omega}^q$ (or $\boldsymbol{\alpha} = \boldsymbol{\sigma}$ in the case $p = \infty$). For a sequence $\boldsymbol{v} \in \mathbb{R}^{\mathbb{N}}$ with $\|\boldsymbol{\sigma}\boldsymbol{v}\|_{\ell^\infty} < \infty$ and $\|\boldsymbol{\omega}\boldsymbol{v}\|_{\ell^q} < \infty$ let $J_n$ be the set of indices corresponding to the $n$ largest elements of the sequence $\boldsymbol{\sigma}|\boldsymbol{v}|$. Then*

$$\|\boldsymbol{v} - P_{J_n}\boldsymbol{v}\|_{\ell^p_{\boldsymbol{\alpha}}} \leq \|P_{J_{n+1}}\tfrac{\boldsymbol{\omega}}{\boldsymbol{\sigma}}\|_{\ell^q}^{-sq}\|\boldsymbol{v}\|_{\ell^q_{\boldsymbol{\omega}}}, \quad s := \tfrac{1}{q} - \tfrac{1}{p}.$$

*For $\boldsymbol{\alpha} \equiv 1$ it holds that $\boldsymbol{\sigma} = \boldsymbol{\omega}^{q/(p-q)}$ and the inequality simplifies to*

$$\|\boldsymbol{v} - P_{J_n}\boldsymbol{v}\|_{\ell^p} \leq \|P_{J_{n+1}}\boldsymbol{\omega}\|_{\ell^{1/s}}^{-1}\|\boldsymbol{v}\|_{\ell^q_{\boldsymbol{\omega}}}, \quad s := \tfrac{1}{q} - \tfrac{1}{p}.$$

**Proof.** We start by proving the assertion for $p = \infty$. Without loss of generality, we can assume that $v$ is ordered such that the sequence $\boldsymbol{\sigma}|\boldsymbol{v}|$ is decreasing. Under this assumption $J_n = [n]$. The choice $p = \infty$ implies $\boldsymbol{\alpha} = \boldsymbol{\sigma}$ and $\|(I - P_{J_n})\boldsymbol{\alpha}\boldsymbol{v}\|_{\ell^p} = \|(I - P_{[n]})\boldsymbol{\sigma}\boldsymbol{v}\|_{\ell^\infty} = \boldsymbol{\sigma}_{n+1}|\boldsymbol{v}_{n+1}|$. Now the bound $\boldsymbol{\sigma}_n|\boldsymbol{v}_n| \leq \|P_{J_n}\tfrac{\boldsymbol{\omega}}{\boldsymbol{\sigma}}\|_{\ell^q}^{-1}\|\boldsymbol{\omega}\boldsymbol{v}\|_{\ell^q}$ follows from

$$\sum_{k=1}^n \Big(\tfrac{\boldsymbol{\omega}_k}{\boldsymbol{\sigma}_k}\Big)^q(\boldsymbol{\sigma}_n|\boldsymbol{v}_n|)^q \leq \sum_{k=1}^n \Big(\tfrac{\boldsymbol{\omega}_k}{\boldsymbol{\sigma}_k}\Big)^q\boldsymbol{\sigma}_k^q|\boldsymbol{v}_k|^q = \sum_{k=1}^n \boldsymbol{\omega}_k^q|\boldsymbol{v}_k|^q \leq \|\boldsymbol{\omega}\boldsymbol{v}\|_{\ell^q}^q.$$

This proves the claim for $p = \infty$. The case $p < \infty$ can be reduced to $p = \infty$ using Hölder's inequality via

$$\|(I - P_{J_n})\boldsymbol{\alpha}\boldsymbol{v}\|_{\ell^p}^p = \|((I - P_{J_n})\boldsymbol{\sigma}\boldsymbol{v})^{p-q}((I - P_{J_n})\boldsymbol{\omega}\boldsymbol{v})^q\|_{\ell^1} \leq \|(I - P_{J_n})\boldsymbol{\sigma}\boldsymbol{v}\|_{\ell^\infty}^{p-q}\|\boldsymbol{\omega}\boldsymbol{v}\|_{\ell^q}^q.$$

The claim follows by using the weighted Stechkin bound for the factor

$$\|(I - P_{J_n})\boldsymbol{\sigma}\boldsymbol{v}\|_{\ell^\infty} \leq \|P_{J_{n+1}}\frac{\boldsymbol{\omega}}{\boldsymbol{\sigma}}\|_{\ell^q}^{-1}\|\boldsymbol{\omega}\boldsymbol{v}\|_{\ell^q}.$$

■

The preceding lemma is a weighted generalisation of Stechkin's Lemma 1.1 with which it coincides for the choice $\boldsymbol{\alpha} = \boldsymbol{\sigma} = \boldsymbol{\omega} \equiv 1$. In this setting, the parameter $q$ has to be chosen as small as possible to

exploit the decay of the sequence $\boldsymbol{v}$ and increase the rate of convergence $s$. When using the weighted Stechkin estimate of Lemma 2.1 this is not necessary since the decay of the sequence can be measured by means of the sequence $\boldsymbol{\omega}$.

To get a better intuition of the derived results, note that each of the sequences $\boldsymbol{\alpha}$, $\boldsymbol{\sigma}$ and $\boldsymbol{\omega}$ controls a different aspect of the estimate. The sequence $\boldsymbol{\alpha}$ determines how the truncation error is measured, $\boldsymbol{\sigma}$ controls the truncation strategy and $\boldsymbol{\omega}$ measures the decay of the sequence. However, due to the constraint $\boldsymbol{\alpha}^p = \boldsymbol{\sigma}^{p-q}\boldsymbol{\omega}^q$ only two of these sequences can be chosen freely. Typically, these are $\boldsymbol{\alpha}$ and $\boldsymbol{\omega}$.

In the remainder of this section, the roles of the different parameters that occur in Lemma 2.1 are discussed with illustrative examples. We start with an examination of $p$ and $\boldsymbol{\alpha}$, which should be chosen to obtain an appropriate error norm $\|\bullet\|_{\ell^p_{\boldsymbol{\alpha}}}$ as in the following four examples.

**Example 2.2** (Sobolev and spectral Barron spaces on the torus)**.** Suppose that $f$ is a function on the 1-torus $\mathbb{T}$ and let $v$ be its sequence of Fourier coefficients. Then $p = 1$ and $p = 2$ together with the weight sequence $\boldsymbol{\alpha}(k)_j := (1+j^2)^{k/2}$ provide a natural choice of parameters since the Sobolev and spectral Barron norms (cf. [15]) of $f$ are then defined by

$$\|f\|_{H^k(\mathbb{T})} := \|\boldsymbol{v}\|_{\ell^2_{\boldsymbol{\alpha}(k)}} \qquad \text{and} \qquad \|f\|_{B^k(\mathbb{T})} := \|\boldsymbol{v}\|_{\ell^1_{\boldsymbol{\alpha}(k)}} \qquad \text{for any } k \in \mathbb{R}.$$

**Example 2.3** (Sobolev and Besov spaces)**.** Consider the Sobolev space $W^{k,p}$ of functions defined on the interval $[0,1]$ equipped with the Lebesgue measure, with $k \geq 1$ and $1 \leq p \leq \infty$. A natural basis for this space is the hierarchical spline basis of degree $k$. It can be shown (see e.g. [3]) that for any $v \in W^{k,p}$,

$$\|v\|_{W^{k,p}} \asymp \|\boldsymbol{v}\|_{\ell^p_{\boldsymbol{\omega}(k)}} \qquad \text{with} \qquad \boldsymbol{\omega}(k)_{\ell,j} := 2^{k\ell}.$$

A simple example for such a basis is provided in Appendix B. These results can be extended to the wider class of Besov spaces $B^k_q(L^p)$ for $0 < p = q \leq \infty$ [3, 37].

**Example 2.4.** Another useful choice of $p$ and $\boldsymbol{\alpha}$ can be made when $f \in L^\infty(\mathcal{X}, \gamma)$ for any measurable set $\mathcal{X}$ and probability measure $\gamma$. Let $\boldsymbol{v}$ be the sequence of coefficients of $f$ with respect to the basis $\{B_j\}_{j\in\mathbb{N}}$ and define the sequence $\boldsymbol{\alpha}_j := \|B_j\|_{L^\infty(\mathcal{X},\gamma)}$. Then, by triangle inequality,

$$\|f\|_{L^p(\mathcal{X},\gamma)} \leq \|f\|_{L^\infty(\mathcal{X},\gamma)} \leq \|\boldsymbol{v}\|_{\ell^1_{\boldsymbol{\alpha}}}.$$

By choosing weights so that $\boldsymbol{\alpha}_j := \|B_j\|_{L^\infty(\mathcal{X},\gamma)} + \|B_j'\|_{L^\infty(\mathcal{X},\gamma)}$, one may also arrive at bounds of the form $\|f\|_{L^\infty(\mathcal{X},\gamma)} + \|f'\|_{L^\infty(\mathcal{X},\gamma)} \leq \|\boldsymbol{f}\|_{\ell^1_{\boldsymbol{\alpha}}}$, reflecting how steeper weights encourage more smoothness (cf. [46]). This bound for instance is used in the proof of Theorem 6.1 in [1], which provides dimension independent convergence rates for unweighted least squares approximation in high dimensions. However, the proof relies on a suboptimal weighted version of Stechkin's lemma, which we discuss further in Section 2.1.

**Example 2.5.** Another useful application of Lemma 2.1 is given by the choice $p = \infty$ and $\boldsymbol{\alpha} \equiv 1$. The choice $p = \infty$ requires $\boldsymbol{\sigma} = \boldsymbol{\alpha} \equiv 1$ and since $\boldsymbol{\alpha} \equiv 1$, the bound simplifies to

$$\|(I - P_{J_n})\boldsymbol{v}\|_{\ell^\infty} \leq \|P_{J_{n+1}}\boldsymbol{\omega}\|^{-1}_{\ell^q}\|\boldsymbol{v}\|_{\ell^q_{\boldsymbol{\omega}}}.$$

Replacing $\boldsymbol{v}$ by the monotonisation

$$\boldsymbol{v}^{\min}_k := \min_{j\leq k}|\boldsymbol{v}_j| \quad \text{for all} \quad k \in \mathbb{N}$$

yields $J_n = [n]$, $\|(I - P_{[n]})\boldsymbol{v}^{\min}\|_{\ell^\infty} = \boldsymbol{v}^{\min}_{n+1}$ and $\|\boldsymbol{v}^{\min}\|_{\ell^q_{\boldsymbol{\omega}}} \leq \|\boldsymbol{v}\|_{\ell^q_{\boldsymbol{\omega}}}$ and simplifies the bound even further to

$$\boldsymbol{v}^{\min}_{n+1} \leq \|P_{[n+1]}\boldsymbol{\omega}\|^{-1}_{\ell^q}\|\boldsymbol{v}\|_{\ell^q_{\boldsymbol{\omega}}}.$$

Next, we examine the parameters $q$ and $\boldsymbol{\omega}$ and illustrate the benefits of the weighted version of Stechkin's lemma in terms of convergence. The subsequent two examples aim to provide an intuition for the choice of $q$ and $\boldsymbol{\omega}$, which should be chosen to capture the asymptotic decay of the sequence $\boldsymbol{v}$ in the reference norm $\|\bullet\|_{\ell^q_{\boldsymbol{\omega}}}$.

**Example 2.6** (The choice of $q$ and $\boldsymbol{\omega}$ for algebraic decay)**.** Consider the algebraically decaying sequence $\boldsymbol{v}_j = j^{-\rho_{\mathrm{alg}}}$ for some $\rho_{\mathrm{alg}} > 1$. To compare Lemma 1.1 and Lemma 2.1, let $\boldsymbol{\alpha} \equiv 1$ and $0 < q < p \leq \infty$ be arbitrary but fixed. Moreover, define $\bar{q} := \frac{1}{q}$ and $\bar{p} := \frac{1}{p}$. Then Lemma A.2 provides the equivalence

$$\|(1 - P_{J_n})\boldsymbol{v}\|_{\ell^p} \sim (n+1)^{-(\rho_{\mathrm{alg}} - \bar{p})}.$$

This rate is a benchmark against which both versions of Stechkin's lemma can be compared. By Lemma A.2 it holds that

$$\frac{\bar{q}}{\rho_{\mathrm{alg}} - \bar{q}} \leq \|\boldsymbol{v}\|^q_{\ell^q} \leq \frac{\rho_{\mathrm{alg}}}{\rho_{\mathrm{alg}} - \bar{q}}$$

for any $\bar{q} \in (\bar{p}, \rho_{\mathrm{alg}})$. Stechkin's lemma thus yields the bound

$$\|(1 - P_{J_n})\boldsymbol{v}\|_{\ell^p} \leq (n+1)^{-s}\|\boldsymbol{v}\|_{\ell^q} \leq (n+1)^{-(\bar{q}-\bar{p})}\left(\frac{\rho_{\mathrm{alg}}}{\rho_{\mathrm{alg}} - \bar{q}}\right)^{\bar{q}}.$$

As $\bar{q}$ approaches the upper bound $\rho_{\mathrm{alg}}$, the predicted rate of convergence approaches the optimal rate $\rho_{\mathrm{alg}} - \bar{p}$. However, at the same time the factor $\|\boldsymbol{v}\|_{\ell^q}$ diverges to infinity. This makes the bound only useful for large $n$ or small $\bar{q}$, i.e. small $s = \bar{q} - \bar{p}$. Although a different weight sequence $\boldsymbol{\omega}$ cannot provide a faster rate of convergence, it can change the asymptotic constant. To this end we define the algebraically increasing sequence $\boldsymbol{\omega}_j = j^r$ for some $r < \rho_{\mathrm{alg}} - \bar{q}$.

This choice implies $\boldsymbol{\sigma}_j = j^{-r/(sp)}$. The technical Lemmas A.1 and A.2 in the appendix yield the bounds

$$\|P_{J_{n+1}}\tfrac{\boldsymbol{\omega}}{\boldsymbol{\sigma}}\|^q_{\ell^q} \geq \frac{(n+1)^{r/s+1}}{r/s+1} \qquad \text{and} \qquad \frac{\bar{q}}{\rho_{\mathrm{alg}} - r - \bar{q}} \leq \|\boldsymbol{v}\|^q_{\ell^q_{\boldsymbol{\omega}}} \leq \frac{\rho_{\mathrm{alg}} - r}{\rho_{\mathrm{alg}} - r - \bar{q}}.$$

Applying Lemma 2.1, we obtain the bound

$$\|(1 - P_{J_n})\boldsymbol{v}\|_{\ell^p} \leq \|P_{J_{n+1}}\tfrac{\boldsymbol{\omega}}{\boldsymbol{\sigma}}\|^{-sq}_{\ell^q}\|\boldsymbol{v}\|_{\ell^q_{\boldsymbol{\omega}}} \leq \frac{(n+1)^{-(r+s)}}{(r/s+1)^{-s}}\left(\frac{\rho_{\mathrm{alg}} - r}{\rho_{\mathrm{alg}} - r - \bar{q}}\right)^{\bar{q}}.$$

As in the unweighted case, the factor $\|\boldsymbol{v}\|_{\ell^q_{\boldsymbol{\omega}}}$ diverges as $r$ increases or $q$ decreases. But in contrast to the unweighted case, we can choose different values for $q$ than in the classical Stechkin estimate while maintaining the same rate of convergence. Denote by $\widetilde{q}$ the value of $\bar{q}$ chosen in the standard Stechkin estimate and recall that $\widetilde{q} < \rho_{\mathrm{alg}}$. We can hence choose $r = \widetilde{q} - \bar{q}$ and take the limit $q \to p$, leading to the estimate

$$\|(1 - P_{J_n})\boldsymbol{v}\|_{\ell^p} \leq (n+1)^{-(\bar{q}-\bar{p})}\left(\frac{\bar{p}}{\rho_{\mathrm{alg}} - \widetilde{q}}\right)^{\bar{p}}.$$

Compared to the unweighted case, the asymptotic constant in the weighted case is significantly smaller. A comparison of these constants is given in Figure 2.1. These constants makes the Stechkin bound viable even for small values of $n$.

**Example 2.7** (The choice of $q$ and $\boldsymbol{\omega}$ for exponential decay)**.** Consider the exponentially decaying sequence $\boldsymbol{v}_j = \rho_{\mathrm{exp}}^{j-1}$ with $\rho_{\mathrm{exp}} \in (0, 1)$. To compare Lemma 1.1 and Lemma 2.1, let $\boldsymbol{\alpha} \equiv 1$ and $0 < q < p \leq \infty$ be arbitrary but fixed. Moreover, define $\bar{q} := \frac{1}{q}$ and $\bar{p} := \frac{1}{p}$. Then

$$\|(1 - P_{J_n})\boldsymbol{v}\|_{\ell^p} = \rho_{\mathrm{exp}}^n(1 - \rho_{\mathrm{exp}}^p)^{-\bar{p}}.$$

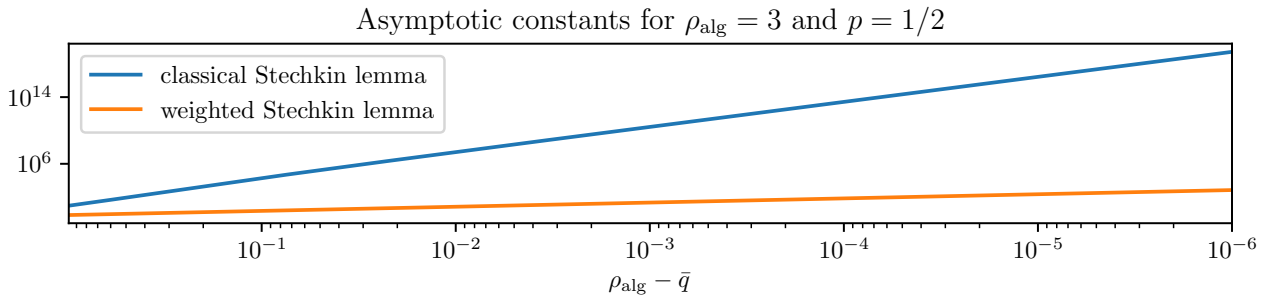


FIGURE 2.1. Behaviour of the asymptotic constants of the classical and weighted Stechkin estimates as the desired rate of convergence $s$ approaches the optimal rate $s^* := \rho_{\mathrm{alg}} - p^{-1}$.

This rate is a benchmark against which both versions of Stechkin's lemma can be compared. The classical Stechkin lemma yields the bound

$$\|(1 - P_{J_n})\boldsymbol{v}\|_{\ell^p} \leq (n+1)^{-s}\|\boldsymbol{v}\|_{\ell^q} = (n+1)^{-s}(1 - \rho_{\exp}^q)^{-\bar{q}}, \quad \text{with } s = \bar{q} - \bar{p}.$$

Notably, even though the factor $\|\boldsymbol{v}\|_{\ell^q}$ does no longer impose a lower limit on $q$, the optimal exponential rate of convergence cannot be recovered. Moreover, the asymptotic constant still grows without bounds when $q$ decreases. This illustrates that Lemma 1.1 cannot fully exploit the decay of the sequence, which renders the estimates only useful as an asymptotic statement or for small values of $s$.

To compare the preceding bound with the weighted bound of Lemma 2.1, we choose the exponentially growing weight sequence $\boldsymbol{\omega}_j = r^{-(j-1)}$ for some $r \in (\rho_{\exp}, 1)$. This choice implies $\boldsymbol{\sigma}_j = r^{(j-1)/(sp)}$ and consequently

$$\|P_{J_{n+1}}\tfrac{\boldsymbol{\omega}}{\boldsymbol{\sigma}}\|_{\ell^q}^q = \frac{r^{-(n+1)/s} - 1}{r^{-1/s} - 1} \geq r^{-n/s} \qquad \text{and} \qquad \|\boldsymbol{v}\|_{\ell^q_{\boldsymbol{\omega}}}^q = \frac{1}{1 - (\rho_{\exp}/r)^q}.$$

Lemma 2.1 then yields

$$\|(1 - P_{J_n})\boldsymbol{v}\|_{\ell^p} \leq \|P_{J_{n+1}}\tfrac{\boldsymbol{\omega}}{\boldsymbol{\sigma}}\|_{\ell^q}^{-sq}\|\boldsymbol{v}\|_{\ell^q_{\boldsymbol{\omega}}} \leq r^n(1 - (\rho_{\exp}/r)^q)^{-\bar{q}}.$$

This shows that in contrast to the classical Stechkin inequality, the weighted Stechkin inequality can actually recover an exponential rate of convergence. This rate is even independent of $q$ but the asymptotic constant grows without bounds when $r$ approaches $\rho_{\exp}$.

### 2.1. **Relation to previous results**

Lemma 2.1 is not the first extension of Stechkin's lemma to the weighted case. Another extension was proposed in [46], which we briefly recall. For a fixed parameter $p$ and sequence $\widetilde{\boldsymbol{\omega}}$ they consider the *weighted $r$-sparse approximation error*

$$\sigma_r(v)_{\ell^p_{\boldsymbol{\alpha}}} := \min_{\substack{I \subseteq \mathbb{N} \\ \widetilde{\boldsymbol{\omega}}(I) \leq r}} \|(1 - P_I)\boldsymbol{v}\|_{\ell^p_{\boldsymbol{\alpha}}}, \qquad \widetilde{\boldsymbol{\omega}} := \boldsymbol{\alpha}^{p/(2-p)},$$

where for any subset $I \subseteq \mathbb{N}$ the weighted cardinality is defined by $\widetilde{\boldsymbol{\omega}}(I) := \sum_{i \in I} \widetilde{\boldsymbol{\omega}}_i^2$. To bound this error, for a given sequence $\boldsymbol{v}$ and a threshold $r > 0$ the set of indices $J_n$ corresponding to the $n$ largest elements of the sequence $\widetilde{\boldsymbol{\omega}}|\boldsymbol{v}|$ is considered. Moreover, let $n(r) := \max\{n \in \mathbb{N} : \widetilde{\boldsymbol{\omega}}(J_n) \leq r\}$. By maximality of $n(r)$, it holds that $\widetilde{\boldsymbol{\omega}}(J_{n(r)+1}) > r$. Consequently,

$$\sigma_r(\boldsymbol{v})_{\ell^p_{\boldsymbol{\alpha}}} \leq \|(1 - P_{J_{n(r)}})\boldsymbol{v}\|_{\ell^p_{\boldsymbol{\alpha}}}. \tag{2.1}$$

Using the unweighted version of Stechkin's lemma, it is concluded in [46, Theorem 3.2] that

$$\sigma_r(\boldsymbol{v})_{\ell^p_{\boldsymbol{\alpha}}} \leq (r - \|\widetilde{\boldsymbol{\omega}}\|^2_{\ell^\infty})^{-s} \|\widetilde{\boldsymbol{\omega}}^{(2-q)/q}\|_{\ell^q}, \quad s := \frac{1}{q} - \frac{1}{p},$$

for all $0 < q < p \leq 2$ and $r > \|\widetilde{\boldsymbol{\omega}}\|^2_{\ell^\infty}$. The main weakness of this statement comes from the requirement $r > \|\widetilde{\boldsymbol{\omega}}\|^2_{\ell^\infty}$. This condition requires that the sequence $\widetilde{\boldsymbol{\omega}}$ is bounded and implies at the same time that the bound only applies for a large threshold $r$, i.e. asymptotically. Hence, either the sparse vectors are part of a finite dimensional space, which contradicts the asymptotical nature of the result, or the sparse vectors are in an infinite dimensional space but the sequence is asymptotically constant, which only results in a very limited generalisation of the classical Stechkin lemma. These shortcomings can be eliminated by using our weighted version of Stechkin's lemma. In fact, applying Lemma 2.1 with $\boldsymbol{\alpha} = \widetilde{\boldsymbol{\omega}}^{(2-p)/p}$ and $\boldsymbol{\omega} = \widetilde{\boldsymbol{\omega}}^{(2-q)/q}$ to the bound (2.1) yields the subsequent corollary of Lemma 2.1.

**Corollary 2.8.** *For $\widetilde{\boldsymbol{\omega}} \in [0,\infty]^{\mathbb{N}}$ and $0 < q < p \leq 2$ define $\boldsymbol{\alpha} := \widetilde{\boldsymbol{\omega}}^{(2-p)/p}$, $\boldsymbol{\sigma} := \widetilde{\boldsymbol{\omega}}^{-1}$ and $\boldsymbol{\omega} := \widetilde{\boldsymbol{\omega}}^{(2-q)/q}$. Let $\boldsymbol{v} \in \ell^\infty_{\boldsymbol{\sigma}} \cap \ell^q_{\boldsymbol{\omega}}$ and let $J_n$ be the set of indices corresponding to the $n$ largest elements of $\boldsymbol{\sigma}|\boldsymbol{v}|$. Then, for any $r \geq 0$,*

$$\boldsymbol{\sigma}_r(\boldsymbol{v})_{\ell^p_{\boldsymbol{\alpha}}} \leq \|(1 - P_{J_{n(r)}})\boldsymbol{v}\|_{\ell^p_{\boldsymbol{\alpha}}} \leq r^{-s}\|\boldsymbol{v}\|_{\ell^q_{\boldsymbol{\omega}}}, \quad s := \frac{1}{q} - \frac{1}{p}.$$

This new weighted version of Stechkin's lemma results in improved bounds in the original work [46] as well as derived works such as [1].

**Example 2.9.** Let $\{B_j\}_{j\in\mathbb{N}}$ be an $L^2(Y,\rho)$-orthonormal basis and identify $v \in L^2(Y,\rho)$ with its sequence of coefficients $\boldsymbol{v} \in \ell^2$. Moreover, define the weight sequence $\boldsymbol{\omega}_j := \|B_j\|_{w,\infty}$ and the model class

$$\mathcal{M} := \{\boldsymbol{v} \in \ell^2 \,:\, \|\boldsymbol{v}\|_{\ell^0_{\boldsymbol{\omega}}} \leq r\}.$$

Combining Proposition 1.4 and Theorem 1.5 with Corollary 2.8 yields the bound

$$\begin{aligned}
\|v - v_{\mathcal{M},\boldsymbol{y}}\| &\leq \|v - v_{\mathcal{M}}\| + \tfrac{2}{\sqrt{1-\delta}}\|v - v_{\mathcal{M}}\|_{w,\infty} \\
&\leq \|\boldsymbol{v} - \boldsymbol{v}_{\mathcal{M}}\|_{\ell^2} + \tfrac{2}{\sqrt{1-\delta}}\|\boldsymbol{v} - \boldsymbol{v}_{\mathcal{M}}\|_{\ell^1_{\boldsymbol{\omega}}} \\
&\leq (1 + \tfrac{2}{\sqrt{1-\delta}})\|\boldsymbol{v} - \boldsymbol{v}_{\mathcal{M}}\|_{\ell^1_{\boldsymbol{\omega}}} \\
&\leq r^{-1}(1 + \tfrac{2}{\sqrt{1-\delta}})\|\boldsymbol{v}\|_{\ell^{1/2}_{\boldsymbol{\omega}^{3/2}}},
\end{aligned}$$

which holds with high probability if $n \gtrsim r\log^3(r)\delta^{-2}$.

## 2.2. The relation of $\ell^p_{\boldsymbol{\omega}}$ to other spaces

It is of general interest to examine the relation of weighted sequence spaces depending on exponents and the weight sequences, which is the topic of this section. We first substantiate our claim that $\boldsymbol{\omega}$ measures the decay of the sequence $\boldsymbol{v}$ by noting that a sequence in $\ell^q_{\boldsymbol{\omega}}$ decays in modulus with a rate of $\boldsymbol{\omega}^{-1}$.

**Lemma 2.10.** *Let $\boldsymbol{v} \in \ell^q_{\boldsymbol{\omega}}$. Then $|\boldsymbol{v}_k| \leq \boldsymbol{\omega}_k^{-1}\|\boldsymbol{v}\|_{\ell^q_{\boldsymbol{\omega}}}$.*

**Proof.** Obviously, $\boldsymbol{\omega}_k|\boldsymbol{v}_k| \leq \|\boldsymbol{\omega}\boldsymbol{v}\|_{\ell^q}$. ∎

The regularity in the $\ell^q_{\boldsymbol{\omega}}$ space is described by the two parameters $q$ and $\boldsymbol{\omega}$. The preceding lemma implies that a sequence $\boldsymbol{v} \in \ell^p_{\widetilde{\boldsymbol{\omega}}}$ also lies in $\ell^q_{\boldsymbol{\omega}}$, if its upper bound lies in this space, i.e.

$$\|\boldsymbol{v}\|_{\ell^q_{\boldsymbol{\omega}}} \leq \|\widetilde{\boldsymbol{\omega}}^{-1}\|\boldsymbol{v}\|_{\ell^p_{\widetilde{\boldsymbol{\omega}}}}\|_{\ell^q_{\boldsymbol{\omega}}} = \|\widetilde{\boldsymbol{\omega}}^{-1}\|_{\ell^q_{\boldsymbol{\omega}}}\|\boldsymbol{v}\|_{\ell^p_{\widetilde{\boldsymbol{\omega}}}}.$$

This indicates that the exponent parameter $q$ can be increased by simultaneously increasing the weight sequence parameter $\boldsymbol{\omega}$. This is made precise in the subsequent lemma.

**Lemma 2.11.** *Let $0 \leq q \leq p \leq \infty$ and $\boldsymbol{\omega}, \widetilde{\boldsymbol{\omega}} \in [0,\infty]^{\mathbb{N}}$. Then for any sequence $\boldsymbol{v} \in \ell^p_{\widetilde{\boldsymbol{\omega}}}$, it holds that*

$$\|\boldsymbol{v}\|_{\ell^q_{\boldsymbol{\omega}}} \leq \|\widetilde{\boldsymbol{\omega}}^{-1}\|_{\ell^{1/s}_{\boldsymbol{\omega}}} \|\boldsymbol{v}\|_{\ell^p_{\widetilde{\boldsymbol{\omega}}}}, \quad s = \tfrac{1}{q} - \tfrac{1}{p}.$$

**Proof.** By Hölder's inequality

$$\|\boldsymbol{v}\|^q_{\ell^q_{\boldsymbol{\omega}}} = \|\boldsymbol{\omega}^q \boldsymbol{v}^q\|_{\ell^1} = \|(\tfrac{\boldsymbol{\omega}}{\widetilde{\boldsymbol{\omega}}})^q (\widetilde{\boldsymbol{\omega}}\boldsymbol{v})^q\|_{\ell^1} \leq \|(\tfrac{\boldsymbol{\omega}}{\widetilde{\boldsymbol{\omega}}})^q\|_{\ell^r} \|(\widetilde{\boldsymbol{\omega}}\boldsymbol{v})^q\|_{\ell^t} = \|\tfrac{\boldsymbol{\omega}}{\widetilde{\boldsymbol{\omega}}}\|^q_{\ell^{rq}} \|\widetilde{\boldsymbol{\omega}}\boldsymbol{v}\|^q_{\ell^{tq}},$$

where $r \in [1,\infty]$ and $t \in [1,\infty]$ satisfy $\frac{1}{r} + \frac{1}{t} = 1$. Choosing $t = \frac{p}{q}$ yields the claim. ∎

**Example 2.12** (Sparse polynomial approximation rates in Gaussian Sobolev spaces)**.** Let $\gamma$ be the standard Gaussian measure on $\mathbb{R}$ and $\{B_j\}_{j\in\mathbb{N}}$ be the basis of normalised Hermite polynomial in $L^2(\mathbb{R},\gamma)$. Since these polynomials constitute an Appell sequence, it holds that

$$\|v\|_{H^k(\mathbb{R},\gamma)} = \|\boldsymbol{v}\|_{\ell^2_{\widetilde{\boldsymbol{\omega}}(k)}} \qquad \text{with} \qquad \widetilde{\boldsymbol{\omega}}(k)_j := \sqrt{\textstyle\sum_{\ell=0}^{\min\{j,k\}} \frac{\Gamma(j+1)}{\Gamma(j-\ell+1)}} \asymp j^{k/2} := \boldsymbol{\omega}(k)_j.$$

Applying Lemma 2.1 yields the following bound for the best $n$-term approximation $v_n$ of $v$:

$$\|v - v_n\|_{L^2(\mathbb{R},\gamma)} = \|\boldsymbol{v} - P_{J_n}\boldsymbol{v}\|_{\ell^2} \leq \|P_{J_{n+1}}\boldsymbol{\omega}(k-\varepsilon)\|^{-1}_{\ell^2} \|\boldsymbol{v}\|_{\ell^1_{\boldsymbol{\omega}(k-\varepsilon)}} \lesssim (n+1)^{-(k+1-\varepsilon)/2} \|\boldsymbol{v}\|_{\ell^1_{\boldsymbol{\omega}(k-\varepsilon)}}.$$

The required weighted summability $\boldsymbol{v} \in \ell^1_{\boldsymbol{\omega}(k-\varepsilon)}$ can usually not be inferred directly from the smoothness of the function. However, since $\|\widetilde{\boldsymbol{\omega}}(k+1)^{-1}\|_{\ell^2_{\boldsymbol{\omega}(k-\varepsilon)}}$ is finite, Lemma 2.11 can be used to obtain the more natural summability condition $\boldsymbol{v} \in \ell^2_{\widetilde{\boldsymbol{\omega}}(k+1)}$. Consequently, for arbitrary $\varepsilon > 0$,

$$\|v - v_n\|_{L^2(\mathbb{R},\gamma)} \lesssim (n+1)^{-(k+1-\varepsilon)/2} \|v\|_{H^{k+1}(\mathbb{R},\gamma)}.$$

**Remark 2.13** (Best $n$-term rates in higher dimensions)**.** To briefly discuss the best $n$-term rates in higher dimensions, we consider isotropic weight sequences of the form $\bar{\boldsymbol{\omega}}(a) = \boldsymbol{\omega}(a)^{\otimes M}$, where $a \in (0,\infty)$ determines growth of $\boldsymbol{\omega}(a)$ (rates for anisotropic product weight sequences should follow by similar arguments). To obtain worst-case rates for the approximation, we apply Lemmas 2.1 and 2.11

$$\|\boldsymbol{u} - P_{J_n\boldsymbol{u}}\boldsymbol{u}\|_{\ell^2} \leq \|P_{J_{n+1}}\bar{\boldsymbol{\omega}}(a)\|^{-1}_{\ell^2} \|\boldsymbol{u}\|_{\ell^1_{\bar{\boldsymbol{\omega}}(a)}} \leq \|P_{J_{n+1}}\bar{\boldsymbol{\omega}}(a)\|^{-1}_{\ell^2} \|\bar{\boldsymbol{\omega}}(A)^{-1}\|_{\ell^2_{\bar{\boldsymbol{\omega}}(a)}} \|\boldsymbol{u}\|_{\ell^2_{\bar{\boldsymbol{\omega}}(A)}}$$

and compute an upper bound for the decay rate $\varepsilon(n) := \|P_{J_{n+1}}\bar{\boldsymbol{\omega}}(a)\|^{-1}_{\ell^2}$. Then, for fixed $a$, we choose the parameter $A > a$ as small as possible while ensuring that $\|\bar{\boldsymbol{\omega}}(A)^{-1}\|_{\ell^2_{\bar{\boldsymbol{\omega}}(a)}}$ is finite.

**Exponential decay (analytic regularity).** Consider the weight sequence $\boldsymbol{\omega}(a)_j = f_a(j)$ with $f_a(x) := \exp(ax)$. In this case, there exists a constant $c_R$ such that

$$\varepsilon(n) \lesssim n^{-(M-1)/(2M)} \exp(-c_R a n^{1/M})$$

and it holds that $\|\bar{\boldsymbol{\omega}}(A)^{-1}\|_{\ell^2_{\bar{\boldsymbol{\omega}}(a)}} < \infty$ for any $a < A$.

**Algebraic decay (mixed Sobolev regularity).** Consider the weight sequence $\boldsymbol{\omega}(a)_j = g_a(j)$ with $g_a(x) := (x+1)^a$. In this case, we obtain the bound

$$\varepsilon(n) \lesssim n^{-(a+1/2)} \ln(n)^{a(M-1)}.$$

and it holds that $\|\bar{\boldsymbol{\omega}}(A)^{-1}\|_{\ell^2_{\bar{\boldsymbol{\omega}}(a)}} < \infty$ for any $a < A - \frac{1}{2}$. Proofs for these statements can be found in appendix C. Note that interestingly, the best $n$-term approximation rate seems to depend on

the dimension $M$ in the exponential case while it is independent of $M$ in the algebraic case. Together with Example 2.12, this provides best $n$-term $L^2$-approximation rates in the Sobolev spaces $H^{k,\mathrm{mix}}(\mathbb{R}^R, \gamma^{\otimes M})$ for the tensor product Hermite polynomial basis. It can also be shown [31] that the similar rate $\|(I - P_{J_n})u\|_{L^2} \leq n^{-k}\|u\|_{H^{k,\mathrm{mix}}}$ also holds (up to logarithmic factors) for the hierarchical tensor product spline basis in the Sobolev spaces $H^{k,\mathrm{mix}}([0,1]^M, \lambda^{\otimes M})$ from Example 2.3.

Finally, we use these insights to highlight the relation of weighted sequences spaces to other wellknwon sequence spaces. We start our discussion with the relation of $\ell^p_{\boldsymbol{\omega}}$ for different values of $p$ and $\boldsymbol{\omega}$. In particular, we show that $\ell^q$ can not be embedded into $\ell^p_{\boldsymbol{\omega}}$ for any $p$ and any unbounded $\boldsymbol{\omega}$. It is clear that this is not possible, because otherwise Lemma 2.10 would provide decay rates for sequences in $\ell^p$. A concrete counterexample is provided in the proof of the subsequent lemma.

**Lemma 2.14.** *Let $0 < q \leq p \leq \infty$ and $\boldsymbol{\omega} \lesssim \widetilde{\boldsymbol{\omega}} \in [0,\infty]^{\mathbb{N}}$. Then*

(i) *$\ell^q_{\boldsymbol{\omega}} \subseteq \ell^p_{\boldsymbol{\omega}}$ and $\ell^p_{\widetilde{\boldsymbol{\omega}}} \subseteq \ell^p_{\boldsymbol{\omega}}$.*

*Moreover,*

(ii) *if $1 \lesssim \boldsymbol{\omega}$ is bounded, then $\ell^p_{\boldsymbol{\omega}} \simeq \ell^p$ and*

(iii) *if $\boldsymbol{\omega}$ is unbounded, then $\ell^q \not\subseteq \ell^p_{\boldsymbol{\omega}} \subseteq \ell^p$ for any $q > 0$.*

**Proof.** The two inclusions $\ell^q_{\boldsymbol{\omega}} \subseteq \ell^p_{\boldsymbol{\omega}}$ and $\ell^p_{\widetilde{\boldsymbol{\omega}}} \subseteq \ell^p_{\boldsymbol{\omega}}$ follow by definition and the assertion (ii) holds because $\|\boldsymbol{v}\|_{\ell^p} \lesssim \|\boldsymbol{\omega}\boldsymbol{v}\|_{\ell^p} \leq \|\boldsymbol{\omega}\|_{\ell^\infty}\|\boldsymbol{v}\|_{\ell^p}$. Hence, the mapping $\boldsymbol{v} \mapsto \boldsymbol{\omega}^{-1}\boldsymbol{v}$ provides an isometry between $\ell^p$ and $\ell^p_{\boldsymbol{\omega}}$. To show the assertion (iii), assume that $\boldsymbol{\omega}$ is unbounded. Then there exists a strictly increasing function $\sigma : \mathbb{N} \to \mathbb{N}$, which defines a subsequence of $\boldsymbol{\omega}$ such that $\boldsymbol{\omega}_{\sigma(k)} \geq 2^k$ for all $k \in \mathbb{N}$. Now let $\varepsilon > 0$ and define the sequence $\boldsymbol{v}$ by

$$\boldsymbol{v}_{\sigma(j)} = j^{-(1+\varepsilon)/q} \quad \text{and} \quad \boldsymbol{v}_k = 0 \quad \text{otherwise.}$$

This sequence satisfies $\boldsymbol{v} \in \ell^q$ and $\boldsymbol{v} \notin \ell^p_{\boldsymbol{\omega}}$ since

$$\|\boldsymbol{v}\|^q_{\ell^q} = \sum_{k\in\mathbb{N}} |\boldsymbol{v}_k|^q = \sum_{j\in\mathbb{N}} |\boldsymbol{v}_{\boldsymbol{\sigma}(j)}|^q = \sum_{j\in\mathbb{N}} j^{-(1+\varepsilon)} < \infty \quad \text{and}$$

$$\|\boldsymbol{v}\|^p_{\ell^p_{\boldsymbol{\omega}}} = \sum_{k\in\mathbb{N}} |\boldsymbol{\omega}_k \boldsymbol{v}_k|^p = \sum_{j\in\mathbb{N}} |\boldsymbol{\omega}_{\sigma(j)}\boldsymbol{v}_{\sigma(j)}|^p \geq \sum_{j\in\mathbb{N}} 2^{jp} j^{-(1+\varepsilon)p/q} = \infty.$$

■

Finally, we examine the relation of the weighted $\ell^q_{\boldsymbol{\omega}}$ spaces to the *monotone $\ell^{q,\mathrm{mon}}$ spaces* (cf. [1]). For any sequence $\boldsymbol{v}$, define the *minimal monotone majorant* $\boldsymbol{v}^{\mathrm{mon}}$ by

$$\boldsymbol{v}^{\mathrm{mon}}_j := \sup_{k\geq j} |\boldsymbol{v}_k| \quad \text{for all} \quad j \in \mathbb{N}.$$

The space $\ell^{q,\mathrm{mon}}$ is then defined as the set of all sequences for which the norm $\|\boldsymbol{v}\|_{\ell^{q,\mathrm{mon}}} := \|\boldsymbol{v}^{\mathrm{mon}}\|_{\ell^q}$ is finite.

**Lemma 2.15.** *Let $0 < p, q \leq \infty$ and $\boldsymbol{\omega} \in [0,\infty]^{\mathbb{N}}$ and define $\varkappa_k := (k+1)^{-1/p}$. Then,*

(i) *$\ell^q_{\boldsymbol{\omega}} \subseteq \ell^{p,\mathrm{mon}}$ if $\boldsymbol{\omega}^{-1} \in \ell^{p,\mathrm{mon}}$ and*

(ii) *$\ell^{p,\mathrm{mon}} \subseteq \ell^q_{\boldsymbol{\omega}}$ if $\boldsymbol{\varkappa} \in \ell^q_{\boldsymbol{\omega}}$.*

**Proof.** For the first assertion, assume that $\boldsymbol{v} \in \ell^q_{\boldsymbol{\omega}}$. Then, by Lemma 2.10, $|\boldsymbol{v}^{\mathrm{mon}}_k| = \sup_{j\geq k}|\boldsymbol{v}_j| \leq \sup_{j\geq k} \boldsymbol{\omega}_j^{-1}\|\boldsymbol{v}\|_{\ell^q_{\boldsymbol{\omega}}} = (\boldsymbol{\omega}^{-1})^{\mathrm{mon}}_k \|\boldsymbol{v}\|_{\ell^q_{\boldsymbol{\omega}}}$. To show the second assertion, let $\boldsymbol{v} \in \ell^{p,\mathrm{mon}}$. Then, by Lemma 1.1, $|\boldsymbol{v}_k| \leq |\boldsymbol{v}^{\mathrm{mon}}_k| \leq \varkappa_k \|\boldsymbol{v}\|_{\ell^{p,\mathrm{mon}}}$ and consequently $\|\boldsymbol{\omega}\boldsymbol{v}\|_{\ell^q} \leq \|\boldsymbol{\omega}\boldsymbol{\varkappa}\|_{\ell^q}\|\boldsymbol{v}\|_{\ell^{p,\mathrm{mon}}}$. ■

## 3. Sparse approximation of parametric PDEs

This section is concerned with an application of the weighted Stechkin lemma for a popular class of functions where weighted sparsity is encountered naturally. In what follows, we consider solutions of parametric PDEs that have become popular in the field of Uncertainty Quantification. We restrict our attention to two prototypical examples mentioned above in Theorems 1.2 and 1.3 that exhibit a holomorphic dependence on the parameter $y$. The proofs of these bounds are typically rather involved and e.g. make use of techniques from complex analysis. With the weighted version of Stechkin's lemma deduced in the preceding section, alternative proofs for such bounds can be derived with more elementary techniques. The principle is demonstrated in this section for the one-dimensional case $L = 1$ as a use case of Lemma 2.1.

Assuming that the coefficient $a(y) \geq \check{a}(y) > 0$ is finite and bounded from below for every $y \in \mathbb{R}$ and $f \in H^{-1}(D)$, Lax–Milgram theorem allows us to define the solution $u(y)$ in the space $H^1_0(D)$ through the variational formulation

$$\int_D a(x,y)\nabla_x u(x,y) \cdot \nabla_x v(x)\,\mathrm{d}x = \int_D f(x)v(x)\,\mathrm{d}x \qquad \text{for all} \qquad v \in H^1_0(D).$$

Moreover, a standard Lax–Milgram a priori estimate tells us that

$$\|u(y)\|_{H^1_0} \leq \check{a}(y)^{-1}\|f\|_{H^{-1}(D)}.$$

Using the machinery of weighted $\ell^p_{\boldsymbol{\omega}}$-spaces developed in Section 2, we now derive a priori best $n$-term convergence bounds for the solution of (1.1) from first principles. Both results rely on the holomorphy of the solution map $y \mapsto u(y) := u(\bullet, y)$ and make use of the following extension of Cauchy's inequality to Banach spaces.

**Theorem 3.1** (Lemma 2.4 in [18])**.** *Let $\rho \in (1,\infty)$, $X$ be a Banach space and $v : B_{\mathbb{C}}(0,\rho) \to X$ be holomorphic such that $\sup_{y\in B_{\mathbb{C}}(0,\rho)}\|v(y)\|_X \leq M < \infty$. Then the power series coefficients $\boldsymbol{v} \in X^{\mathbb{N}}$ of $v$ satisfy*

$$\|\boldsymbol{v}_k\|_X \leq M\rho^{-k}.$$

### 3.1. Affine coefficients

We first consider the model problem (1.1) with affine coefficients (1.2). Summability of the power series of solution $u$ can be shown based on its holomorphy.

**Theorem 3.2.** *Let $\rho > 1$ and the* uniform ellipticity assumption *(UEA)*

$$C := \inf_{x\in D} a_0(x) - \rho\sum_{j\geq 1}|a_j(x)| > 0$$

*be satisfied. Moreover, let $u$ be the solution of the diffusion equation* (1.1) *with affine coefficients* (1.2)*. Then the map $y \mapsto u(y)$ is holomorphic from $[-1,1]$ to $H^1_0(D)$ and belongs to $L^p([-1,1],\gamma;H^1_0(D))$ for all $p \in \mathbb{N}\cup\{\infty\}$, where $\gamma$ is the uniform measure. Moreover, for any $r \in (0,\rho)$, the power series coefficients $\boldsymbol{u}$ of $u$ satisfy the bound*

$$\|\boldsymbol{u}_k\|_{H^1_0(D)} \leq \|f\|_{H^{-1}(D)}C^{-1}r^{-k}.$$

We omit the proof of this theorem since it follows by the same arguments as the one of the more interesting log-affine case in Theorem 3.7. Moreover, we note that in higher dimensions an anisotropic choice of $\rho$ can be used to reduce the regularity assumptions on $a$ as it is done in the log-affine case.

The preceding lemma guarantees a decay of the power series coefficients of $u$. However, in numerical applications an expansion in terms of an orthonormal basis is preferable. For the diffusion equation (1.1) with affine coefficients (1.2), a suitable basis is given by the Legendre polynomials. The

subsequent two lemmas show how the decay of the power series coefficients translates into a decay of the Legendre coefficients.

**Lemma 3.3** (see [36])**.** *Let $v$ satisfy the conditions of Theorem 3.1 and let $\boldsymbol{v}$ be the power series coefficients of $v$. Then*

$$v(z) = \sum_{m\in\mathbb{N}} \widehat{\boldsymbol{v}}_m L_m(z),$$

*where $L_m$ is the $m^{th}$ normalised Legendre polynomial and $\widehat{\boldsymbol{v}}_m$ satisfies*

$$\widehat{\boldsymbol{v}}_m = \sum_{n\in\mathbb{N}} \frac{(2m+1)^{1/2}(m+2n)!}{2^{m+2n}n!(\frac{3}{2})_{m+n}} \boldsymbol{v}_{m+2n},$$

*where the Pochhammer symbol $(a)_k$ is defined as $(a)_0 = 1$, $(a)_{k+1} = (a)_k(a+k)$.*

**Theorem 3.4.** *Let $u$ be the solution of the diffusion equation (1.1) with affine coefficients (1.2). Moreover, let the parameters $\rho > 1$ and $C > 0$ be defined as in Theorem 3.2. Then, for any $r \in (1,\rho)$, the Legendre basis coefficients $\widehat{\boldsymbol{u}}$ of $u$ satisfy*

$$\|\widehat{\boldsymbol{u}}_m\|_{H^1_0(D)} \leq \|f\|_{H^{-1}(D)} C^{-1} \sqrt{2m+1} \frac{r^{-m}}{r^2-1}.$$

*This implies that $\widehat{\boldsymbol{u}} \in \ell^q_{\boldsymbol{\omega}}$ for all $q \in (0,\infty]$ and $\boldsymbol{\omega}_n := p(n)r^n$ with $r \in (1,\rho)$ and any positive function $p$ growing at most polynomially.*

**Proof.** Denote by $\boldsymbol{u}$ the power series coefficients of $u$ and define the double sequence

$$\boldsymbol{\alpha}_{m,k} := \sqrt{2m+1} \frac{k!}{2^k((k-m)/2)!(\frac{3}{2})_{(k+m)/2}}.$$

By Lemma 3.3 it holds that $\widehat{\boldsymbol{u}}_m = \sum_{k\in m+2\mathbb{N}} \boldsymbol{\alpha}_{m,k}\boldsymbol{u}_k$ and hence

$$\|\widehat{\boldsymbol{u}}_m\|_{H^1_0(D)} \leq \|P_{m+2\mathbb{N}}\boldsymbol{u}\|_{\ell^1_{\boldsymbol{\alpha}_m}}, \tag{3.1}$$

where the $\|\boldsymbol{u}\|_{\ell^1_{\alpha m}} = \sum_{k\in\mathbb{N}} \alpha_{m,k}\|\boldsymbol{u}_k\|_{H^1_0(D)}$. This bound is tight, since equality holds for $\boldsymbol{u}_{k+1} = \boldsymbol{v}_k\boldsymbol{u}_k$ and any non-negative sequence $0 \leq \boldsymbol{v} \in \ell^1_{\alpha_m}$. By Theorem 3.2 it holds that $\|\boldsymbol{u}_k\|_{H^1_0(D)} \leq cr^{-k}$ with $c := \|f\|_{H^{-1}(D)}C^{-1}$. Moreover, expressing the Pochhammer symbol in terms of the Gamma function yields the bound

$$(\tfrac{3}{2})_k = \frac{\Gamma(k+\frac{3}{2})}{\Gamma(\frac{3}{2})} > \underbrace{\left(\min_{z\in[0,\infty)} \frac{\Gamma(z+\frac{3}{2})}{\Gamma(z+1)}\right)}_{=\Gamma(\frac{3}{2})} \frac{\Gamma(k+1)}{\Gamma(\frac{3}{2})} = k!\ .$$

Substituting both bounds into (3.1) yields

$$\|\widehat{\boldsymbol{u}}_m\|_{H^1_0(D)} \leq c\sqrt{2m+1}r^{-m} \sum_{n\in\mathbb{N}} \underbrace{2^{-(m+2n)}\binom{m+2n}{n}}_{=p_{m+2n}(n)\leq 1} r^{-2n} \leq \frac{c\sqrt{2m+1}r^{-m}}{r^2-1},$$

where $p_{m+2n}$ is the probability mass function of a binomial distribution with $m+2n$ trials and a success probability of $\frac{1}{2}$. The final claim follows directly from this bound and the ratio test for series convergence. ∎

**Remark 3.5.** It is possible to use Lemma 2.1 to bound (3.1). But as shown in Examples 2.6 and 2.7, Stechkin's lemma cannot fully exploit the decay of a weight sequence. In fact, if possible it is preferable to derive a bound directly as in the previous proof.

**Example 3.6.** Let $u$ be the solution of the diffusion equation (1.1) with affine coefficients (1.2). We denote by $L_m$ the $m^{\text{th}}$ normalised Legendre polynomial and by $\widehat{\boldsymbol{u}}$ the Legendre basis coefficients of $u$. Moreover, define the weight sequence $\boldsymbol{\omega}_j := \|L_j\|_\infty = \sqrt{2j+1}$ and the model class

$$\mathcal{M} := \{\boldsymbol{v} \in \ell^2 \,:\, \|\boldsymbol{v}\|_{\ell^0_{\boldsymbol{\omega}}} \leq r\}.$$

We know from Example 2.9 that

$$\|u - u_{\mathcal{M},\boldsymbol{y}}\| \leq r^{-1}(1 + \tfrac{2}{\sqrt{1-\delta}})\|\boldsymbol{u}\|_{\ell^{1/2}_{\boldsymbol{\omega}^{3/2}}}$$

holds with high probability if $n \gtrsim r\log^3(r)\delta^{-2}$. Theorem 3.4 guarantees that

$$\|\boldsymbol{u}\|^{1/2}_{\ell^{1/2}_{\boldsymbol{\omega}^{3/2}}} \lesssim \sum_{m\in\mathbb{N}} (2m+1)^{3/4}\Big(\sqrt{2m+2}r^{-m}\Big)^{1/2} = \sum_{m\in\mathbb{N}} (2m+1)r^{-m/2} = \frac{\sqrt{r}(\sqrt{r}+1)}{(\sqrt{r}-1)^2}$$

is indeed finite. Note that we can probably obtain better rates by using a faster growing weight sequence $\boldsymbol{\omega}$ and Lemma 2.1 instead of Corollary 2.8.

### 3.2. **Log-affine coefficients**

The analysis of (1.1) with log-affine coefficient (1.3) is much more involved from a theoretical and practical side than the affine case. As in the affine case, we begin by showing holomorphy of the solution $u$. The analysis is based on the approach in [17], where only the affine case was considered and an explicit decay of the coefficients $\|\boldsymbol{u}_\nu\|_{H^1_0(D)}$ was not shown.

**Theorem 3.7.** *For every $x \in D$, let $\boldsymbol{a}(x)$ denote the sequence of coefficients $(\boldsymbol{a}_j(x))_{j\in\mathbb{N}}$. Assume that there exists a sequence $\boldsymbol{\rho} \in (0,\infty)^{\mathbb{N}}$ such that*

$$\sup_{x\in D}\|\boldsymbol{a}(x)\|_{\ell^\infty_{\boldsymbol{\rho}}} = C_1 < \infty \qquad \text{and} \qquad \|\boldsymbol{\rho}^{-1}\|_{\ell^2} = C_2 < \infty.$$

*Then the map $y \mapsto u(y)$ is entire and belongs to $L^p(\mathbb{R},\gamma;H^1_0(D))$ for all $p \in \mathbb{N}$, where $\gamma$ denotes the Gaussian measure. Moreover, the power series coefficients satisfy the bound*

$$\|\boldsymbol{u}_{\boldsymbol{\nu}}\|_{H^1_0(D)} \leq \|f\|_{H^{-1}(D)}\exp(\|\boldsymbol{\nu}\|_1)\Big(\frac{\boldsymbol{\nu}\boldsymbol{\rho}}{C_1}\Big)^{-\boldsymbol{\nu}}.$$

**Proof.** We start by providing a lower bound for $\breve{a}(y) > 0$. Since

$$\begin{aligned}\inf_x \exp((\boldsymbol{a}(x),y)_{\ell^2}) &= \exp(\inf_x(\boldsymbol{a}(x)\boldsymbol{\rho},\boldsymbol{\rho}^{-1}y)_{\ell^2})\\ &\geq \exp(-\sup_x\|\boldsymbol{a}(x)\|_{\ell^\infty_{\boldsymbol{\rho}}}\|\boldsymbol{\rho}^{-1}y\|_{\ell^1})\\ &= \exp(-C_1\|\boldsymbol{\rho}^{-1}y\|_{\ell^1}),\end{aligned}$$

it holds that $\breve{a}(y) \geq \exp(-C_1\|\boldsymbol{\rho}^{-1}y\|_{\ell^1})$. The integrability of $u$ now follows by the simple calculation

$$\begin{aligned}\mathbb{E}_y[\breve{a}(y)^{-k}] &\leq \mathbb{E}_y[\exp(kC_1\|\boldsymbol{\rho}^{-1}y\|_{\ell^1})]\\ &= \prod_{j\in\mathbb{N}}\mathbb{E}_{y_j}[\exp(kC_1\boldsymbol{\rho}_j^{-1}|y_j|)]\\ &\lesssim \prod_{j\in\mathbb{N}}\exp(\tfrac{1}{2}k^2C_1^2\boldsymbol{\rho}_j^{-2})\\ &= \exp(\tfrac{1}{2}k^2C_1^2C_2^2)\\ &< \infty.\end{aligned}$$

We now show that the extension of the map $y \mapsto u(y)$ to the complex domain is analytic. Following [17], we start by defining for $\boldsymbol{r} \in (0,\infty)^{\mathbb{N}}$ the open polydiscs $U_{\boldsymbol{r}} := \prod_{k\in\mathbb{N}} B(0,\boldsymbol{r}_k)$ on which $u$ is uniformly bounded by

$$\|u\|_{L^\infty(U_{\boldsymbol{r}};H^1_0(D))} \leq \exp(-C_1\|\boldsymbol{\rho}^{-1}\boldsymbol{r}\|_{\ell^1})\|f\|_{H^{-1}(D)}. \tag{3.2}$$

Now, we introduce for any coefficient field $a$ the operator $B(a) : v \mapsto -\operatorname{div}_x(a\nabla_x v)$ mapping from $H^1_0(D)$ to $H^{-1}(D)$ and decompose the map $y \mapsto u(y)$ into the chain of holomorphic maps

$$y \mapsto a(y) \mapsto B(a(y)) \mapsto B(a(y))^{-1} \mapsto B(a(y))^{-1}f = u(y).$$

The first map is holomorphic by definition and the second and last map are continuous linear maps and thereby also holomorphic. The third map is the operator inversion which is holomorphic at any invertible $B$. Since $B(a(y))$ is invertible for every $y \in \mathbb{C}^{\mathbb{N}}$, the map $y \mapsto u(y)$ is entire. Applying Cauchy's inequality in Theorem 3.1 to (3.2), we hence obtain

$$\|\boldsymbol{u}_{\boldsymbol{\nu}}\|_{H^1_0(D)} \leq \|f\|_{H^{-1}(D)} \exp(-C_1\|\boldsymbol{\rho}^{-1}\boldsymbol{r}\|_{\ell^1})\boldsymbol{r}^{-\boldsymbol{\nu}}.$$

Choosing for every fixed multi-index $\boldsymbol{\nu}$ the sequence $\boldsymbol{r}_k := \frac{\boldsymbol{\nu}_k\boldsymbol{\rho}_k}{C_1}$ yields $\|\boldsymbol{\rho}^{-1}\boldsymbol{r}\|_{\ell^1} = \frac{1}{C_1}\|\boldsymbol{\nu}\|_{\ell^1}$ and proves the result. ∎

In the setting of log-affine coefficients (1.3), a suitable basis is given by the Hermite polynomials. Similar to the Lemma 3.3 and Theorem 3.4, the subsequent two results show how the decay of the power series coefficients translates into a decay of the Hermite coefficients.

**Lemma 3.8.** *Let $v$ satisfy the conditions of Theorem 3.1 and let $\boldsymbol{v}$ be the power series coefficents of $v$. Then*

$$v(z) = \sum_{m\in\mathbb{N}} \widehat{\boldsymbol{v}}_m H_m(z),$$

*where $H_m$ is the $m^{th}$ normalised Hermite polynomial and $\widehat{\boldsymbol{v}}_m$ satisfies*

$$\widehat{\boldsymbol{v}}_m = \sum_{n\in\mathbb{N}} \frac{(m+2n)!}{2^n n!\sqrt{m!}} \boldsymbol{v}_{m+2n}.$$

**Proof.** For every $m \in \mathbb{N}$, let $He_m$ denote the $k^{\text{th}}$ monic probabilist's Hermite polynomial. Then, by [44, Chapter 11, Section 110],

$$z^k = \sum_{\substack{n\in\mathbb{N}\\ 2n\leq k}} \frac{k!}{2^n n!(k-2n)!} He_{k-2n}(z).$$

Plugging this into the power series expansion for $v$ yields

$$v(z) = \sum_{\substack{k,n\in\mathbb{N}\\ 2n\leq k}} \frac{k!}{2^n n!(k-2n)!}\boldsymbol{v}_k He_{k-2n}(z) = \sum_{\substack{k,n,m\in\mathbb{N}\\ 2n\leq k\\ m=k-2n}} \frac{k!}{2^n n!m!}\boldsymbol{v}_k He_m(z) = \sum_{n,m\in\mathbb{N}} \frac{(m+2n)!}{2^n n!m!}\boldsymbol{v}_{m+2n} He_m(z).$$

Substituting $He_m = \sqrt{m!}H_m$ yields the desired relation. ∎

**Theorem 3.9.** *Let $u$ be the solution of the diffusion equation (1.1) with log-affine coefficients (1.3). Moreover, let the parameters $\boldsymbol{\rho}$ and $C_1$ be defined as in Theorem 3.7. Finally, let $L = 1$ and assume that $\boldsymbol{\rho} > C_1$. Then, the Hermite basis coefficients $\widehat{\boldsymbol{u}}$ of $u$ satisfy $\|\widehat{\boldsymbol{u}}_m\|_{H^1_0(D)} \lesssim \frac{\|f\|_{H^{-1}(D)}}{\sqrt{(m-1)!}}$. This implies that $\widehat{\boldsymbol{u}} \in \ell^q_{\boldsymbol{\omega}}$ for all $q \in (0,\infty]$ and $\boldsymbol{\omega}$ that grow subfactorially.*

**Proof.** Define the double sequence

$$\boldsymbol{\alpha}_{m,k} := \frac{k!}{2^{(k-m)/2}((k-m)/2)!\sqrt{m!}}.$$

By Lemma 3.8 it holds that $\widehat{\boldsymbol{u}}_m = \sum_{k\in m+2\mathbb{N}} \boldsymbol{\alpha}_{m,k}\boldsymbol{u}_k$. Hence

$$|\widehat{\boldsymbol{u}}_m| \leq \|P_{m+2\mathbb{N}}\boldsymbol{u}\|_{\ell^1_{\boldsymbol{\alpha}_m}} \leq \|(1-P_{[m-1]})\boldsymbol{u}\|_{\ell^1_{\boldsymbol{\alpha}_m}}. \tag{3.3}$$

This bound is tight, since equality holds for any sequence $\boldsymbol{u} \geq 0$ with $\boldsymbol{u}_k = 0$ for all $k \in m+2\mathbb{N}$. By Theorem 3.7 it holds that $\|\boldsymbol{u}_k\|_{H^1_0(D)} \leq \|f\|_{H^{-1}(D)} \exp(k)(\frac{k\rho}{C_1})^{-k}$ and by Stirling's approximation

$$\sqrt{2\pi k}(\tfrac{k}{e})^k \exp(\tfrac{1}{12k+1}) < k! < \sqrt{2\pi k}(\tfrac{k}{e})^k \exp(\tfrac{1}{12k}).$$

Substituting all three estimates into (3.3) and assuming $m \geq 1$ yields

$$\begin{aligned}\|\widehat{\boldsymbol{u}}_m\|_{H^1_0(D)} &\leq \|f\|_{H^{-1}(D)}\sqrt{\frac{2(m+1)}{m!}}\sum_{k\geq m}\left(\frac{k-m}{e}\right)^{-(k-m)/2}\left(\frac{\boldsymbol{\rho}}{C_1}\right)^{-k}\\ &\leq 2\|f\|_{H^{-1}(D)}\left(\sqrt{e}+\tfrac{e}{2}+\frac{(C_1/\boldsymbol{\rho})^{m+3}}{1-C_1/\boldsymbol{\rho}}\right)\frac{1}{\sqrt{(m-1)!}}.\end{aligned}$$

■

**Corollary 3.10.** *Let $\widehat{\boldsymbol{u}}$ be the sequences of Hermite basis coefficients from Theorem 3.9 with $L = 1$ and assume that $\rho \geq C_1$. Then, if $\boldsymbol{\omega}_j = \widetilde{r}^{j/2}$ with $\widetilde{r} \in [1,2)$ and $J_n$ is the set of indices corresponding to the $n$ largest elements of the sequence $\boldsymbol{\omega}_k^{-1}\|\widehat{\boldsymbol{u}}_{\boldsymbol{k}}\|_{H^1_0(D)}$, it holds that*

$$\|(1-P_{J_n})\widehat{u}\|_{L^2} \lesssim \|f\|_{H^{-1}(D)}\frac{\sqrt{2}}{\sqrt{2}-\sqrt{\widetilde{r}}}\widetilde{r}^{-n/2}.$$

**Proof.** From Theorem 3.9, we know that $\|\widehat{\boldsymbol{u}}_m\|_{H^1_0(D)} \lesssim \frac{\|f\|_{H^{-1}(D)}}{\sqrt{(m-1)!}}$. Since $m! \geq 2^{m-1}$, it holds that $\frac{1}{\sqrt{(m-1)!}} \lesssim 2^{-m/2} =: \boldsymbol{\gamma}_m$. Applying Lemma 2.1 with $p = 2$, $q = 1$ and $\boldsymbol{\alpha} \equiv 1$ yields

$$\|(1-P_{J_n})\widehat{\boldsymbol{u}}\|_{L^2} = \|(1-P_{J_n})\widehat{\boldsymbol{u}}\|_{\ell^2} \leq \|P_{J_{n+1}}\boldsymbol{\omega}^2\|_{\ell^1}^{-1/2}\|\widehat{\boldsymbol{u}}\|_{\ell^1_{\boldsymbol{\omega}}} \lesssim \|f\|_{H^{-1}(D)}\|P_{J_{n+1}}\boldsymbol{\omega}^2\|_{\ell^1}^{-1/2}\|\boldsymbol{\gamma}\|_{\ell^1_{\boldsymbol{\omega}}}.$$

The claim follows since $\|\boldsymbol{\gamma}\|_{\ell^1_{\boldsymbol{\omega}}} = \frac{\sqrt{2}}{\sqrt{2}-\sqrt{\widetilde{r}}}$ and $\|P_{J_{n+1}}\boldsymbol{\omega}^2\|_{\ell^1} \geq \|P_{[n+1]}\boldsymbol{\omega}^2\|_{\ell^1} \geq \widetilde{r}^n$. ■

**Remark 3.11.** Note that the proofs of Theorem 3.4 and 3.9 rely essentially on the formulas in Lemma 3.3 and Lemma 3.8. The similarity of these formulas indicates a deeper relation stemming from the explicit representations of $(p_n, L_m)_{L^2}$ and $(p_n, H_m)_{L^2}$ with $p_n(z) := z^n$. We conjecture that similar representations can be derived for all families of orthonormal polynomials by means of the corresponding three-term recurrence relation.

## 4. Sparse approximation using tensor trains

In this section we consider sparse approximation problems in a high-dimensional setting where weighted sparse vectors can be identified with tensors. We show that tensors in $\ell^q_{\boldsymbol{\omega}}$ can be approximated efficiently in a model class of tensor trains with (weighted) sparse component tensors. The derivation of this relies heavily on results in [38], from which we recall some theorems. For the sake of completeness and since the proofs foster some interesting insights, they are also provided.

Finally, we provide a practical algorithm to obtain these representations which provides an alternative for classical sparse approximation algorithms (such as weighted $\ell^1$-minimisation) that circumvents the CoD.

### 4.1. **Tensor train representation of sparse tensors**

This section recalls basic representation results for sparse tensors that are originally due to [38]. We first introduce some basic operations on tensors.

**Definition 4.1** (Vectorisation)**.** For any tensor $A \in \mathbb{R}^{d_1 \times \cdots \times d_M}$ of order $M \in \mathbb{N}_{>0}$, the vectorisation of $A$ is a vector $\mathrm{vec}(A) \in \mathbb{R}^{d_1 \cdots d_M}$ defined by the equality

$$A_{i_1,i_2,\ldots,i_d} = \mathrm{vec}(A)_{\sum_{k\in[M]} i_k D_k} \qquad \text{with} \qquad D_k := \prod_{\ell=k+1}^{M} d_\ell \qquad \text{and} \qquad D_M := 1.$$

**Definition 4.2** (Unfolding [43])**.** For any tensor $A \in \mathbb{R}^{d_1 \times \cdots \times d_M}$ and $k \in [M]$, the $k$-unfolding of $A$ is a matrix $\mathrm{unfold}_k(A) \in \mathbb{R}^{C_1 \times C_2}$ with $C_1 = \prod_{j=1}^{k} d_j$ and $C_2 = \prod_{j=k+1}^{M} d_j$, defined by the equality $\mathrm{vec}(A) = \mathrm{vec}(\mathrm{unfold}_k(A))$.

**Definition 4.3** (Orthogonality)**.** A tensor $A \in \mathbb{R}^{d_1 \times \cdots \times d_M}$ is called left-orthogonal, if

$$\mathrm{unfold}_{M-1}(A)^\intercal \,\mathrm{unfold}_{M-1}(A) = I.$$

It is called right-orthogonal if

$$\mathrm{unfold}_1(A)\,\mathrm{unfold}_1(A)^\intercal = I.$$

**Definition 4.4** (Contraction)**.** Given two tensors $A \in \mathbb{R}^{d_1 \times \cdots \times d_M}$ and $B \in \mathbb{R}^{d_M \times \cdots \times d_N}$, we define the *contraction* of $A$ and $B$ along the last dimension of $A$ and the first dimension of $B$ as

$$(A \circ B)_{i_1,\ldots,i_{M-1},i_{M+1},\ldots,i_N} := \sum_{i_M \in [d_M]} A_{i_1,\ldots,i_M} B_{i_M,\ldots,i_N}.$$

The *tensor train* (TT) decomposition [43] represents a tensor of order $M$ as the contraction of $M$ lower order tensors. A tensor $A \in \mathbb{R}^{d_1 \times \cdots \times d_M}$ is said to have a TT representation of *rank* $r \in \mathbb{N}^{M-1}$ if

$$A = A^{(1)} \circ \cdots \circ A^{(M)}$$

with *component tensors* $A^{(k)} \in \mathbb{R}^{r_{k-1} \times d_k \times r_k}$ and the convention that $r_{-1} = r_M = 1$. By fixing the second index of every $A^{(k)}$ to $i_k$, we obtain a matrix $A^{(k)}_{i_k}$. The entries of $A$ can then be computed by

$$A_{i_1,\ldots,i_M} = A^{(1)}_{i_1} \cdots A^{(M)}_{i_M}.$$

Now suppose that the tensor $A$ is $R$-sparse, i.e. that there exists a set $J$ of size $R$ such that $A_i \neq 0$ if and only if $i = (i_1, \ldots, i_M) \in J$. Then $A$ can be represented as the sum of $R$ rank-1 tensors,

$$A = \sum_{i \in J} e_{i_1} \otimes \cdots \otimes (A_i e_{i_k}) \otimes \cdots \otimes e_{i_M}, \tag{4.1}$$

where $e_{i_l} \in \mathbb{R}^{d_l}$ are the standard basis vectors and the choice of the index $k \in [M]$ is arbitrary. Since every summand is a TT of rank 1, the sum (4.1) can be represented as a TT of rank $R$.

**Lemma 4.5** (Section 4.1 in [43])**.** *Let $A, B \in \mathbb{R}^{d_1 \times \cdots \times d_M}$ be two tensors given in TT format*

$$A_i = A^{(1)}_{i_1} \cdots A^{(M)}_{i_M}, \qquad B_i = B^{(1)}_{i_1} \cdots B^{(M)}_{i_M}.$$

*The sum $C = A + B$ can be represented in TT format with components*

$$C^{(1)}_{i_1} = \begin{bmatrix} A^{(1)}_{i_1} & B^{(1)}_{i_1} \end{bmatrix}, \qquad C^{(k)}_{i_k} = \begin{bmatrix} A^{(k)}_{i_k} & \\ & B^{(k)}_{i_k} \end{bmatrix}, \qquad C^{(M)}_{i_M} = \begin{bmatrix} A^{(M)}_{i_M} \\ B^{(M)}_{i_M} \end{bmatrix},$$

*where $k = 2, \ldots, M-1$ and empty spaces denote blocks of zeros of appropriate dimension.*

The proof of Lemma 4.5 follows directly from the definition of the TT decomposition. Together with the decomposition (4.1) it implies that any $R$-sparse tensor $A \in \mathbb{R}^{d_1 \times \cdots \times d_M}$ can be represented as a TT of rank $R$.

This decomposition can be written as

$$A = P^{(1)} \circ \cdots \circ P^{(k-1)} \circ C \circ P^{(k+1)} \circ \cdots \circ P^{(M)},$$

with $P^{(1)} \in \{0,1\}^{1 \times d_1 \times R}$, $P^{(j)} \in \{0,1\}^{R \times d_j \times R}$ for $1 < j < M$, $P^{(M)} \in \{0,1\}^{R \times d_M \times 1}$, and $C \in \mathbb{R}^{R \times d_k \times R}$. If $J = \{i^1, \ldots, i^R\}$, then by the definition of the component tensors in Lemma 4.5 it holds that

$$\operatorname{unfold}_2(P^{(1)}) = \begin{bmatrix} e_{i_1^1} & \cdots & e_{i_1^R} \end{bmatrix}, \quad \operatorname{unfold}_2(P^{(j)}) = \begin{bmatrix} e_{i_j^1} & & \\ & \ddots & \\ & & e_{i_j^R} \end{bmatrix}, \quad \operatorname{unfold}_1(P^{(M)}) = \begin{bmatrix} e_{i_M^1}^\intercal \\ \vdots \\ e_{i_M^R}^\intercal \end{bmatrix},$$

where $C$ exhibits the same sparsity pattern as the corresponding $P^{(j)}$. Note that all components in this representation are $R$-sparse and that similar representations exist for $k = 1$ or $k = d$.

Now consider the case that $i_1^1 = i_1^2 = k$. Then the column $e_k$ appears at least twice in the matricisation $\operatorname{unfold}_2(P^{(1)})$ resulting in an ambiguous representation of the tensor. This is a principal effect of the representation in Lemma 4.5 and is not specific to the sparse TT decomposition. In classical tensor algorithms, uniqueness of the representation is restored (up to orthogonal transformations) by performing a sequence of rank-revealing QR decompositions on the factors $P^{(j)}$. However, since the QR decomposition is not guaranteed to preserve sparsity, we introduce a sparse QC decomposition $X = QC$, where $Q$ is orthogonal and sparse and $C$ is sparse. The idea behind this decomposition is that the image space of $X$ is spanned by those standard basis vectors $e_i$ for which the row vector $e_i^\intercal X$ is non-zero. We can hence define $Q$ as the sparse orthogonal matrix containing these standard basis vectors as its columns.

To rigorously define this decomposition, recall that any $R$-sparse matrix $A \in \mathbb{R}^{n \times m}$ can be represented by the three $R$-tuples

$$\operatorname{row}(A) \in [n]^R, \qquad \operatorname{col}(A) \in [m]^R \qquad \text{and} \qquad \operatorname{data}(A) \in \mathbb{R}^R.$$

Here $\operatorname{row}(A)_i$ and $\operatorname{col}(A)_i$ are the row and column indices of the $i^{\text{th}}$ non-zero entry in $A$ and $\operatorname{data}(A)_i$ is its value. Conversely, given three $R$-tuples $r \in [n]^R$, $c \in [m]^R$ and $d \in \mathbb{R}^R$ such that the pairs $\{(r_i, c_i)\}_{i \in [R]}$ are unique, we can uniquely define an $R$-sparse matrix $\operatorname{coo}(r, c, d)$ with

$$\operatorname{row}(\operatorname{coo}(r,c,d)) = r, \qquad \operatorname{col}(\operatorname{coo}(r,c,d)) = c \qquad \text{and} \qquad \operatorname{data}(\operatorname{coo}(r,c,d)) = d.$$

Finally, define for every $R \in \mathbb{N}$ the $R$-tuples

$$\operatorname{range}(R) := (1, \ldots, R) \qquad \text{and} \qquad \operatorname{ones}(R) := (1, \ldots, 1)$$

as well as the tuple $\operatorname{unique}(r)$ for every tuple $r \in \mathbb{N}^R$, containing only the unique elements of $r$. As usual, we define for any vector $x \in \mathbb{R}^d$ the dimension $\dim(x) := d$.

**Definition 4.6.** Let $X \in \mathbb{R}^{n \times m}$ be an $R$-sparse matrix. Then the *sparse QC decomposition* $X = QC$ is given by

$$Q := \operatorname{coo}(s, \operatorname{range}(r), \operatorname{ones}(r)) \in \mathbb{R}^{n \times r} \qquad \text{and} \qquad C := Q^\intercal X,$$

where $s := \operatorname{unique}(\operatorname{row}(X))$ and $r := \dim(s) \leq R$.

**Lemma 4.7.** *Let $X \in \mathbb{R}^{n \times m}$ be an $R$-sparse matrix and $X = QC$ be its sparse QC decomposition. Then $Q \in \mathbb{R}^{n \times r}$ is orthogonal and $r$-sparse with $r \leq R$ and $C \in \mathbb{R}^{r \times m}$ is $R$-sparse.*

**Proof.** Recall that $Q := \mathrm{coo}(s, \mathrm{range}(r), \mathrm{ones}(r))$ with $s := \mathrm{unique}(\mathrm{row}(X))$ and $r := \dim(s)$. This means that $Q$ is $r$-sparse with $r = \dim(s) \leq \dim(\mathrm{row}(X)) = R$. Moreover, since the $k^{\text{th}}$ column of $Q$ is the standard basis vector $e_{s_k}$ and since the indices in $s$ are unique, it follows that $Q$ is orthogonal. For the same reason, $C = Q^\intercal X$ is just a version of $X$ with the non-zero rows removed. Therefore, $C$ is $R$-sparse. ∎

Applying the sparse QC decomposition sequentially to the unfoldings of all component tensors results in a TT representation

$$A = U^{(1)} \circ \cdots \circ U^{(k-1)} \circ C \circ V^{(k+1)} \circ \cdots \circ V^{(M)}. \tag{4.2}$$

An implementation of this procedure is listed in Algorithm 1. The resulting component tensors $U^{(j)} \in \{0,1\}^{r_{j-1} \times d \times r_j}$ are $r_j$-sparse and left-orthogonal and the component tensors $V^{(j)} \in \{0,1\}^{r_{j-1} \times d \times r_j}$ are $r_{j-1}$-sparse and right-orthogonal. The ranks $r_j$ are uniformly bounded by $R$ and the *core tensor* $C$ remains $R$-sparse. These properties are summarised in the following definition.

**Definition 4.8.** A tensor train representation

$$A = U^{(1)} \circ \cdots \circ U^{(k-1)} \circ C \circ V^{(k+1)} \circ \cdots \circ V^{(M)}$$

is called *sparsely canonicalised with core position $k$* if

(1) $U^{(j)} \in \{0,1\}^{r_{j-1} \times d \times r_j}$ are left-orthogonal and $r_j$-sparse for all $1 \leq j < k$,

(2) $V^{(j)} \in \{0,1\}^{r_{j-1} \times d \times r_j}$ are right-orthogonal and $r_{j-1}$-sparse for all $k < j \leq M$ and

(3) $C \in \mathbb{R}^{r_{k-1} \times d \times r_k}$ is $\min\{r_{k-1}, r_k\}$-sparse.

**Algorithm 1:** Sparse canonicalisation

**input:** Tensor train representation $A = A^{(1)} \circ \cdots \circ A^{(M)}$, desired core position $k$.
**output:** Sparsely canonicalised representation of $A$ with core position $k$.

Initialise $C^{(0)} := I$.
**for** $j = 1$ **to** $k - 1$ **do**
  Define $X^{(j)} := C^{(j-1)} \,\mathrm{unfold}_2(A^{(j)})$.
  Compute the sparse QC decomposition (cf. Definition 4.6) $X^{(j)} = Q^{(j)} C^{(j)}$.
  Define $\mathrm{unfold}_2(U^{(j)}) := Q^{(j)}$.
**end**
Initialise $C^{(M+1)} := I$.
**for** $j = M$ **to** $k + 1$ **do**
  Define $X^{(j)} := \mathrm{unfold}_1(A^{(j)}) C^{(j+1)}$.
  Compute the sparse QC decomposition (cf. Definition 4.6) $(X^{(j)})^\intercal = (Q^{(j)})^\intercal (C^{(j)})^\intercal$.
  Define $\mathrm{unfold}_1(V^{(j)}) := Q^{(j)}$.
**end**
Define $C := C^{(k-1)} A^{(k)} C^{(k+1)}$.
**return** $A = U^{(1)} \circ \cdots \circ U^{(k-1)} \circ C \circ V^{(k+1)} \circ \cdots \circ V^{(M)}$.

### 4.2. **Approximation results**

The deliberations of the preceding section give rise to a model class of tensor trains with sparse component tensors. Moreover, due to the special structure of the component tensors any weighted summability condition on the full tensor translates into a weighted summability condition on the core tensor. This is made precise in the subsequent theorem.

**Theorem 4.9.** *Every $R$-sparse tensor $A \in \mathbb{R}^{d_1 \times \cdots \times d_M}$ can be represented in a sparsely canonicalised TT format*

$$A = U^{(1)} \circ \cdots \circ U^{(k-1)} \circ C \circ V^{(k+1)} \circ \cdots \circ V^{(M)}$$

*with ranks that are uniformly bounded by $R$, independent of the chosen core position $k \in [M]$. Moreover, we can define the operator $Q \in \mathcal{L}(\mathbb{R}^{r_k \times d_k \times r_{k+1}}, \mathbb{R}^{d_1 \times \cdots \times d_M}) \simeq \mathbb{R}^{d_1 \cdots d_M \times r_k d_k r_{k+1}}$ via*

$$Q = \operatorname{unfold}_k(U^{(1)} \circ \cdots \circ U^{(k-1)}) \otimes I_{d_k} \otimes \operatorname{unfold}_1(V^{(k+1)} \circ \cdots \circ V^{(M)})^\intercal, \tag{4.3}$$

*where $\otimes$ denotes the matrix Kronecker product. This means that $A = QC$. Interpreted as a matrix, $Q$ is left-orthogonal and its columns are standard basis vectors and for every $q \in [0, \infty]$ and $\boldsymbol{\omega} \in [0, \infty]^{d_1 \times \cdots \times d_M}$ it holds that*

$$\|A\|_{\ell^q_{\boldsymbol{\omega}}} = \|C\|_{\ell^q_{\boldsymbol{\beta}}},$$

*where $\boldsymbol{\beta} := Q^\intercal \boldsymbol{\omega}$ is a (reshaped) subsequence of $\boldsymbol{\omega}$.*

**Proof.** $Q$ is a linear operator mapping a component tensor from $\mathbb{R}^{r_k \times d_k \times r_{k+1}}$ to the space of full tensors $\mathbb{R}^{d_1 \times \cdots \times d_M}$. After vectorising these tensor spaces, we can interpret $Q$ as a matrix $Q \in \{0,1\}^{d_1 \cdots d_M \times r_k d_k r_{k+1}}$. We now show that $Q$ is an orthogonal matrix where every column is a standard product basis vector. We begin by showing that the matrices

$$B_k := \operatorname{unfold}_{k+1}(U^{(1)} \circ \cdots \circ U^{(k)})$$

are left-orthogonal with columns that are standard basis vectors. Following the lines of [57, Appendix B], this can be proved by induction. For $k = 1$ the assertion is true by construction of $U^{(1)}$. For $k > 1$ it holds that $B_k = (I_{r_{k-1}} \boxtimes B_{k-1}) \operatorname{unfold}_2(U^{(k)})$, where $I_{r_{k-1}} \boxtimes B_{k-1}$ denotes the Kronecker product. The two matrices $(I_{r_{k-1}} \boxtimes B_{k-1})$ and $\operatorname{unfold}_2(U^{(k)})$ are left-orthogonal and their columns are standard basis vectors. This implies the assertion, since the matrix product preserves these properties. A similar argument shows that the matrices

$$D_k := \operatorname{unfold}_1(U^{(k)} \circ \cdots \circ U^{(M)})^\intercal$$

are left-orthogonal with columns that are standard basis vectors. This proves the claim, since $Q = B_{k-1} \otimes I_{d_k} \otimes D_{k+1}$. ∎

Let $A$ be an $R$-sparse coefficient tensor with $\|A\|_{\ell^0_{\boldsymbol{\omega}}} \leq r$. Then Theorem 4.9 ensures that $A = QC$ with

$$\begin{aligned} Q \in \mathcal{Q}_{R,k} := \{ &\operatorname{unfold}_k(U^{(1)} \circ \cdots \circ U^{(k-1)}) \otimes I_{d_k} \otimes \operatorname{unfold}_1(V^{(k+1)} \circ \cdots \circ V^{(M)})^\intercal \\ &: U^{(j)}, V^{(j)} \in \{0,1\}^{r_{j-1} \times d_j \times r_j} \text{ with } r_{j-1}, r_j \leq R \\ &: \text{all } U^{(j)} \text{ are left-orthogonal and } r_j\text{-sparse} \\ &: \text{all } V^{(j)} \text{ are right-orthogonal and } r_{j-1}\text{-sparse} \} \end{aligned}$$

and

$$C \in \mathcal{C}_{Q,r,\boldsymbol{\omega}} := B_{\ell^0_{Q^\intercal \boldsymbol{\omega}}}(0, r) = \{C \in \mathbb{R}^{r_{k-1} \times d_k \times r_k} \,:\, \|C\|_{\ell^0_{Q^\intercal \boldsymbol{\omega}}} \leq r\}.$$

Since such a representation exists for all $k = 1, \ldots, M$, this motivates the definition of the model class

$$\mathcal{M}_{R,r,\boldsymbol{\omega}} := \bigcap_{k \in [M]} \bigcup_{Q \in \mathcal{Q}_{R,k}} Q\mathcal{C}_{Q,r,\boldsymbol{\omega}}.$$

Since $Q^\intercal\boldsymbol{\omega}$ is a subsequence of $\boldsymbol{\omega}$, $\boldsymbol{\omega} \geq 1$ implies $\|C\|_{\ell^0_{Q^\intercal\boldsymbol{\omega}}} \geq \|C\|_{\ell^0}$. The memory footprint of $v \in \mathcal{M}_{R,r,\boldsymbol{\omega}}$ for $\boldsymbol{\omega} \geq 1$ is thus bounded by $(M-1)R + r$. As before, we identify the set of coefficient tensors $\mathcal{M}_{R,r,\boldsymbol{\omega}}$ with the corresponding set of functions. The following corollary then translates the result of Corollary 2.8 to the model class $\mathcal{M}_{R,r,\boldsymbol{\omega}}$ and shows that it exhibits similar approximation rates as the more classical sets of weighted sparsity.

**Corollary 4.10.** *Let $\boldsymbol{\tau} \in ([0,\infty]^{\mathbb{N}})^M$ and $T(R) := \min\{\|P_J\boldsymbol{\tau}\|^2_{\ell^2} : |J| = R+1\}$ be the sum of the $R+1$ smallest elements in $\boldsymbol{\tau}^2$. Moreover, let $0 < q < p \leq 2$ and define $\boldsymbol{\alpha} := \boldsymbol{\tau}^{(2-p)/p}$ and $\boldsymbol{\omega} := \boldsymbol{\tau}^{(2-q)/q}$. Then every $\boldsymbol{v} \in \ell^q_{\boldsymbol{\omega}}(\mathbb{N}^M)$ can be approximated by a tensor $\widetilde{\boldsymbol{v}} \in \mathcal{M}_{R,r,\boldsymbol{\omega}}$ with accuracy*

$$\|\boldsymbol{v} - \widetilde{\boldsymbol{v}}\|_{\ell^p_{\boldsymbol{\alpha}}} \leq \min\{T(R), r\}^{-s}\|\boldsymbol{v}\|_{\ell^q_{\boldsymbol{\omega}}}, \qquad s := \tfrac{1}{q} - \tfrac{1}{p}.$$

**Proof.** Let $J_n$ be defined as in Corollary 2.8 and recall that $\boldsymbol{v}$ can be approximated by the $n$-sparse tensor $\widetilde{\boldsymbol{v}} := P_{J_n}\boldsymbol{v}$ with an error of at most

$$\|(1-P_{J_n})\boldsymbol{v}\|_{\ell^p_{\boldsymbol{\alpha}}} \leq \left(\|P_{J_{n+1}}\boldsymbol{\sigma}^{-1}\|^q_{\ell^q_{\boldsymbol{\omega}}}\right)^{-s}\|\boldsymbol{v}\|_{\ell^q_{\boldsymbol{\omega}}} = \left(\|P_{J_{n+1}}\boldsymbol{\tau}\|^2_{\ell^2}\right)^{-s}\|\boldsymbol{v}\|_{\ell^q_{\boldsymbol{\omega}}}.$$

Define $n(r) := \max\{n \in \mathbb{N} : \|P_{J_n}\boldsymbol{\tau}\|^2_{\ell^2} \leq r\}$ and choose $n = \min\{R, n(r)\}$. Then $\widetilde{\boldsymbol{v}}$ is $R$-sparse and $\boldsymbol{\tau}$-weighted $r$-sparse and Theorem 4.9 guarantees that it can be represented in $\mathcal{M}_{R,r,\boldsymbol{\omega}}$. The desired error bounds follows by case distinction. If $n = R$ then $\|P_{J_{n+1}}\boldsymbol{\tau}\|^2_{\ell^2} \geq T(R)$ by definition of $T(R)$. Hence $(\|P_{J_{n+1}}\boldsymbol{\tau}\|^2_{\ell^2})^{-s} \leq T(R)^{-s}$. If $n = n(r)$ then $\|P_{J_{n+1}}\boldsymbol{\tau}\|^2_{\ell^2} \geq r$ by maximality of $n(r)$. Hence $(\|P_{J_{n+1}}\boldsymbol{\tau}\|^2_{\ell^2})^{-s} \leq r^{-s}$. ∎

Using the model class $\mathcal{M}_{R,r,\boldsymbol{\omega}}$, the optimisation (1.5) becomes feasible on product basis. We use the remainder of this section to provide theoretical guarantees for this optimisation. To apply Proposition 1.4, we first show that the model class satisfies the required nestedness property $\mathcal{M}_{R,r,\boldsymbol{\omega}} - \mathcal{M}_{R,r,\boldsymbol{\omega}} \subseteq \mathcal{M}_{2R,2r,\boldsymbol{\omega}}$ and then show $\mathrm{RIP}_{\mathcal{M}_{R,r,\boldsymbol{\omega}}}(\delta)$ holds with high probability. For this to make sense, we let $b : Y \to \mathbb{R}^d$ be a vector of $L^2(Y,\rho)$-orthonormal basis functions, define the tensor product basis $B(y) := b(y_1) \otimes \cdots \otimes b(y_M)$ and suppose that the weight sequence $\boldsymbol{\omega}$ satisfies $\boldsymbol{\omega}_j \geq \|B_j\|_{L^\infty}$. We now identity the space of coefficents $v \in \mathcal{M}_{R,r,\boldsymbol{\omega}}$ with the corresponding space of functions $y \mapsto (B(y), v)_{\ell^2}$.

**Proposition 4.11.** *It holds that $\mathcal{M}_{R,r,\boldsymbol{\omega}} - \mathcal{M}_{R,r,\boldsymbol{\omega}} \subseteq \mathcal{M}_{2R,2r,\boldsymbol{\omega}}$.*

**Proof.** Let $A, B \in \mathcal{M}_{R,r,\boldsymbol{\omega}}$ with core position $k$. By Lemma 4.5, the difference $C = A - B$ can be represented in TT format with components

$$C^{(1)}_{i_1} = \begin{bmatrix} A^{(1)}_{i_1} & -B^{(1)}_{i_1} \end{bmatrix}, \qquad C^{(j)}_{i_j} = \begin{bmatrix} A^{(j)}_{i_j} & \\ & B^{(j)}_{i_j} \end{bmatrix}, \qquad C^{(M)}_{i_M} = \begin{bmatrix} A^{(M)}_{i_M} \\ B^{(M)}_{i_M} \end{bmatrix},$$

for $j = 2, \ldots, M-1$. After removing duplicate columns in the matricisations of $C^{(j)}$ for $j \neq k$, the resulting tensor satisfies the unweighted sparsity, orthogonality and rank constraints of $\mathcal{M}_{2R,2r,\boldsymbol{\omega}}$. And since

$$\|C^{(k)}\|_{\ell^0_{Q^\intercal\boldsymbol{\omega}}} = \|A - B\|_{\ell^0_{\boldsymbol{\omega}}} \leq \|A\|_{\ell^0_{\boldsymbol{\omega}}} + \|B\|_{\ell^0_{\boldsymbol{\omega}}} \leq 2r,$$

the weighted sparsity constraints are satisfied as well. ∎

The following immediate consequence of Theorem 1.5 provides a bound for the probability of $\mathrm{RIP}_{\mathcal{M}_{R,r,\boldsymbol{\omega}}}(\delta)$.

**Corollary 4.12.** *Fix parameters $\delta, \gamma \in (0,1)$. Let $\{B_j\}_{j\in[D]}$ be orthonormal with respect to the measure $\rho$ and let $w \geq 0$ be any weight function satisfying $\|w^{-1}\|_{L^1} = 1$. Assume the weight sequence satisfies*

$\boldsymbol{\omega}_j \geq \|w^{1/2} B_j\|_{L^\infty}$ *and fix*

$$n \geq C\delta^{-2} r \max\{\log^3(r)\log(d^M), -\log(\gamma)\}.$$

*Let* $y_1, \dots, y_n$ *be drawn independently from* $w^{-1}\rho$. *Then the probability of* $\mathrm{RIP}_{\mathcal{M}_{R,r,\boldsymbol{\omega}}}(\delta)$ *exceeds* $1-\gamma$.

**Proof.** Theorem 4.9 guarantees that every $A \in \mathcal{M}_{R,r,\boldsymbol{\omega}}$ can be written as $A = QC$ with $\|A\|_{\ell^0_{\boldsymbol{\omega}}} = \|C\|_{\ell^0_{\boldsymbol{\beta}}} \leq r$ and where $\boldsymbol{\beta} = Q^\intercal \boldsymbol{\omega}$. This implies that $\mathcal{M}_{R,r,\boldsymbol{\omega}} \subseteq B_{\ell^0_{\boldsymbol{\omega}}}(0,r)$. The assertion follows, since Theorem 1.5 implies $\mathrm{RIP}_{B_{\ell^0_{\boldsymbol{\omega}}}(0,r)}(\delta)$ and, consequently, $\mathrm{RIP}_{\mathcal{M}_{R,r,\boldsymbol{\omega}}}(\delta)$. ∎

### 4.3. **Numerical method**

Recall that we want to solve the optimisation problem (1.5). This means we want to approximate $u \in V = L^2(Y,\rho)$. To present an efficient numerical implementation, we assume that $Y = \mathbb{R}^M$, $\rho = \rho_1^{\otimes M}$ is a product measure and $b : \mathbb{R} \to \mathbb{R}^d$ is a vector of $L^2(\mathbb{R},\rho_1)$-orthonormal basis functions. Defining the product basis

$$B_{i_1\dots i_M}(y_1,\dots,y_M) := b_{i_1}(y_1)\cdots b_{i_M}(y_M),$$

an approximation to $u$ is given by

$$u(y) \approx (v, B(y))_{\mathrm{Fro}} := \sum_{i_1=1}^{d} \cdots \sum_{i_M=1}^{d} v_{i_1\dots i_M} B_{i_1\dots i_M}(y) \tag{4.4}$$

for some $v \in (\mathbb{R}^d)^{\otimes M}$. The preceding section suggests that we can restrict the minimisation in (1.5) to the model class $\mathcal{M}$ of functions of the form (4.4) with $v \in \mathcal{M}_{R,r,\boldsymbol{\omega}}$. To write the resulting optimisation problem compactly, we define the vector $F \in \mathbb{R}^n$ and bounded linear operator $E : (\mathbb{R}^d)^{\otimes M} \to \mathbb{R}^n$ by

$$F_i = \sqrt{w(y^i)}u(y^i) \qquad \text{and} \qquad (Ev)_i = \sqrt{w(y^i)}(v, B(y^i))_{\mathrm{Fro}}. \tag{4.5}$$

Then equation (1.5) is equivalent to the optimisation problem

$$\operatorname*{minimise}_{v \in \mathcal{M}_{R,r,\boldsymbol{\omega}}} \|F - Ev\|_{\ell^2}^2. \tag{4.6}$$

We propose to solve this problem by a sparse variant of the *alternating least squares* (ALS) algorithm introduced in [33, 43]. The ALS method solves (4.6) by refining an initial guess in a sequence of *microsteps*, each optimising a single component tensor while keeping the others fixed. Since every $v \in \mathcal{M}_{R,r,\boldsymbol{\omega}}$ can be written as a sparsely canonicalised tensor train $v = QC$ with core position $k$, the microstep optimising the $k^{\text{th}}$ component tensor can be written as

$$\operatorname*{minimise}_{C \in \mathcal{C}_{Q,r,\boldsymbol{\omega}}} \|F - EQC\|_{\ell^2}^2.$$

The operator $Q$ can be efficiently computed by Algorithm 1[1] and the resulting sparse tensor train representation allows for an efficient evaluation of $EQ$. A classical approach to handle the weighted sparsity constraints in $\mathcal{C}_{Q,r,\boldsymbol{\omega}} = \{C \in \mathbb{R}^{r_{k-1}\times d_k \times r_k} : \|C\|_{\ell^0_{Q^\intercal\boldsymbol{\omega}}} \leq r\}$ is to promote the weighted $\ell^0$-constraints via a weighted $\ell^1$-regularisation term. Using the Hadamard product $(\boldsymbol{\beta}_Q \odot C)_{ijk} := (\boldsymbol{\beta}_Q)_{ijk} C_{ijk}$, the resulting problem reads

$$\operatorname*{minimise}_{C \in \mathbb{R}^{r_{k-1}\times d_k \times r_k}} \|F - EQC\|_{\ell^2}^2 + \lambda \|\boldsymbol{\beta}_Q \odot C\|_1, \tag{4.7}$$

[1] Indeed the operator $Q$ at core position $k$ can be efficiently updated from its value at position $k-1$ or $k+1$ by means of a single sparse QC decomposition.

**Algorithm 2:** Sparse Alternating Least Squares (SALS)

**input:** Data pairs $(y^i, u(y^i)) \in \mathbb{R}^M \times \mathbb{R}$ for $i = 1, \dots, n$, univariate basis functions $\{b_1, \dots, b_d\}$, and weight sequences $\omega_m \in \mathbb{R}^d$ for $m = 1, \dots, M$.

**output:** A coefficient tensor $\boldsymbol{v} \in \mathcal{M}$ such that $y \mapsto (\boldsymbol{v}, B(y))_{\mathrm{Fro}}$ approximates the data.

```
Initialize the coefficient tensor v.
while not converged do
    for k = 1 to M do
        Compute the sparse canonicalisation (4.2) with core position k.
        Compute Q as in (4.3) and β_Q := Q^⊤ ω.
        Update C by solving equation (4.8) and use cross-validation to select λ.
    end
end
return v.
```

where $\boldsymbol{\beta}_Q := Q^\intercal \boldsymbol{\omega}$ can be efficiently computed due to the tensor train representation and sparsity structure of $Q$. Substituting $D = \boldsymbol{\beta}_Q \odot C$ into (4.7) we obtain the standard LASSO problem [22, 48]

$$\underset{D \in \mathbb{R}^{r_{k-1} \times d_k \times r_k}}{\text{minimise}} \ \|F - EQ(\boldsymbol{\beta}_Q^{-1} \odot D)\|_{\ell^2}^2 + \lambda \|D\|_1. \tag{4.8}$$

The regularisation parameter $\lambda$ controls the sparsity of $C$ and must be chosen appropriately to remain in the model class $\mathcal{M}_{R,r,\boldsymbol{\omega}}$. To do this, recall the following two facts.

(1) Theorem 4.9 implies $\|C\|_{\ell^0_{\boldsymbol{\beta}_Q}} = \|\boldsymbol{v}\|_{\ell^0_{\boldsymbol{\omega}}}$, which ensures that the weighted sparsity constraint is satisfied for all components as soon as it is satisfied for the optimised component.

(2) Theorem 4.9 implies $\|C\|_{\ell^0} = \|\boldsymbol{v}\|_{\ell^0}$, which ensures that the rank $R$ is bounded by the number of nonzero entries of the core tensor $\|C\|_{\ell^0}$.

It is thus sufficient to choose $\lambda$ such that $\|C\|_{\ell^0_{\boldsymbol{\beta}_Q}} \leq r$ and $\|C\|_{\ell^0} \leq R$ to remain in $\mathcal{M}_{R,r,\boldsymbol{\omega}}$ during optimisation. Although this would be easy to implement, we propose to choose $\lambda$ by 10-fold cross-validation instead. This allows the algorithm to choose a different regularisation parameter $\lambda$, i.e. a different sparsity level, for every core position $k$. Moreover, since the rank $R$ in the sparsely canonicalised representation depends on the sparsity of the solution of the microstep, the resulting algorithm is inherently rank-adaptive. We call this algorithm *sparse alternating least-squares* (SALS) since it modifies a standard ALS method to work on sparse tensors. A listing of the complete algorithm, in pseudo-code, is provided in Algorithm 2. There it can be seen that the algorithm differs from a standard ALS only in two points.

(1) The standard regression in the microstep is replaced by a (weighted) LASSO.

(2) The rank revealing QR decomposition, commonly used to compute the operators $Q$, is replaced by a sparse QC decomposition.

It is therefore straight-forward to implement.

## 5. Low-rank and sparse Tensor Train approximation

Despite its straightforward sample bound and built-in rank adaptivity, the sparse tensor train model class from the previous section and the associated Algorithm 2 are not optimal, since the resulting

tensor representation does not have minimal rank in general. Motivated by promising practical results with low-rank tensor reconstructions for holomorphic functions as considered in [23, 53], this section introduces a new tensor train format which incorporates sparsity and low-rank.

To illustrate the advantage of this new format, we consider the approximation of the rank-1 function $x \mapsto \exp(x_1 + \ldots + x_M)$ by Legendre polynomials in Appendix D. The remainder of this section is devoted to investigating this idea in the general setting.

### 5.1. **Approximation results**

To obtain an operator $Q$ which still allows for a meaningful concept of sparsity in the component tensor $C$, we replace the sparse QC decomposition from the preceding section with an $\boldsymbol{\omega}$-weighted QC decomposition.

**Definition 5.1.** We say that a matrix $Q$ is *$\boldsymbol{\omega}$-orthogonal* if $Q^\intercal \operatorname{diag}(\boldsymbol{\omega}) Q$ is diagonal.

**Definition 5.2.** Let $A \in \mathbb{R}^{n \times m}$ be a rank-$r$ matrix. A *$\boldsymbol{\omega}$-orthogonal QC decomposition of $A$* is a decomposition $A = QC$ with $Q \in \mathbb{R}^{n \times r}$ and $C \in \mathbb{R}^{r \times m}$ for which $Q$ is orthogonal and $\boldsymbol{\omega}$-orthogonal, i.e. $Q^\intercal Q = I$, and $Q^\intercal \operatorname{diag}(\boldsymbol{\omega}) Q$ is diagonal.

Even though this new decomposition may not retain the sparsity as well as the sparse QC decomposition did, the resulting factors still exhibit a considerable amount of sparsity.

**Lemma 5.3.** *Let $A \in \mathbb{R}^{n \times m}$ be a rank-$r$ matrix. Then there exists an $\boldsymbol{\omega}$-orthogonal QC-decomposition $A = QC$. This decomposition is unique up to reordering of the columns of $Q$. Moreover, if $A$ is $R$-sparse then $r \leq R$ and $Q$ and $C$ are $Rr$-sparse. (Note that the complexity is independent of $n$ and $m$.)*

**Proof.** Let $A = Q_1 C_1$ be the sparse QC decomposition of $A$ and let $C_1 = Q_2 C_2$ be the QR decomposition of $C_1$. Moreover, let $Q_{12} := Q_1 Q_2$ and $U \Lambda U^\intercal$ be the spectral decomposition of $Q_{12}^\intercal \operatorname{diag}(\boldsymbol{\omega}) Q_{12}$, define $Q := Q_{12} U$ and $C := U^\intercal C_2$. Then $A = QC$ by construction and it holds that

$$Q^\intercal \operatorname{diag}(\boldsymbol{\omega}) Q = U^\intercal (Q_{12}^\intercal \operatorname{diag}(\boldsymbol{\omega}) Q_{12}) U = U^\intercal (U \Lambda U^\intercal) U = \Lambda.$$

Note that $Q$ is a product of three orthogonal matrices and hence orthogonal. Since the QR decompostion $A = Q_{12} C_2$ is unique and since the spectral decomposition of $Q_{12}^\intercal \operatorname{diag}(\boldsymbol{\omega}) Q_{12}$ is unique up to reordering of the columns $U$, the matrix $Q = Q_{12} U$ is unique up to reordering of its columns. Now suppose that $A$ is $R$-sparse. Then $Q_1 \in \{0,1\}^{n \times \widetilde{R}}$ with $\widetilde{R} \leq R$. This means that $C_1 \in \mathbb{R}^{\widetilde{R} \times m}$, which yields the standard bound $r \leq \min\{\widetilde{R}, m\} \leq R$. Moreover, since the columns of $Q_1$ are standard basis vectors, only the rows in $\operatorname{row}(Q_1) \in [n]^{\widetilde{R}}$ are nonzero. Consequently, only the same rows can be nonzero in the product $Q = Q_1 (Q_2 U)$. In the worst case, all of the $r$ columns of $Q$ become nonzero for every of these $\widetilde{R}$ rows. This yields a total of $\widetilde{R} r \leq Rr$ nonzero entries. To obtain a sparsity bound for $C$, let $A^\intercal = \widetilde{Q} \widetilde{C}$ be the sparse QC decomposition and $\widetilde{C}^\intercal = Q \bar{C}$ be the $\boldsymbol{\omega}$-weighted QC decomposition. Now define $C = \bar{C} \widetilde{Q}^\intercal$ and observe that $A = QC$ is a valid $\boldsymbol{\omega}$-weighted QC decomposition. Since the rows of $\widetilde{Q}^\intercal$ are standard basis vectors, only the columns in $\operatorname{col}(\widetilde{Q}) \in [m]^{\widetilde{R}}$ are nonzero. Consequently, only the same columns can be nonzero in the product $C = \bar{C} \widetilde{Q}^\intercal$. In the worst case, all of the $r$ rows of $C$ will be nonzero for every of these $\widetilde{R}$ columns. This yields a total of $\widetilde{R} r \leq Rr$ nonzero entries. ∎

Applying the $\boldsymbol{\omega}$-weighted QC decomposition sequentially to the unfoldings of all component tensors results in a TT representation

$$A = QC = U^{(1)} \circ \cdots \circ U^{(k-1)} \circ C \circ V^{(k+1)} \circ \cdots \circ V^{(M)},$$

where the component tensors $U^{(j)} \in \mathbb{R}^{r_{j-1} \times d \times r_j}$ are $r_j R$-sparse and left-orthogonal and the component tensors $V^{(j)} \in \mathbb{R}^{r_{j-1} \times d \times r_j}$ are $r_{j-1} R$-sparse and right-orthogonal, the ranks $r_j$ are uniformly bounded

by $R$ and the *core tensor* $C$ remains $R$-sparse. An implementation of this procedure can be obtained from Algorithm 1 by replacing all sparse QC decompositions with $\boldsymbol{\omega}$-weighted QC decompositions. Analogously to the model class $\mathcal{M}_{R,r,\boldsymbol{\omega}}$, which is based on the sparse QC decomposition, we define the model class

$$\widetilde{\mathcal{M}}_{R,r,\boldsymbol{\omega}} := \bigcap_{k\in[M]} \bigcup_{Q\in\widetilde{\mathcal{Q}}_{k,R,\boldsymbol{\omega}^3}} Q\mathcal{C}_{Q,r,\boldsymbol{\omega}^3},$$

which is based on the $\boldsymbol{\omega}$-weighted QC decomposition. The elements of this new model class are tensors $A = QC$ with $Q \in \widetilde{\mathcal{Q}}_{k,R,\boldsymbol{\omega}}$ and $C \in \mathcal{C}_{Q,r,\boldsymbol{\omega}}$, where

$$\widetilde{\mathcal{Q}}_{k,R,\boldsymbol{\omega}} := \Big\{ Q \in \mathcal{L}(\mathbb{R}^{r_k\times d_k\times r_{k+1}}, \mathbb{R}^{d_1\times\cdots\times d_M}) \;\Big|\; Q \text{ is orthogonal and } \boldsymbol{\omega}^2\text{-orthogonal and } r_k, r_{k+1} \le R \Big\}$$

and

$$\mathcal{C}_{Q,r,\boldsymbol{\omega}} := \{C \in \mathbb{R}^{r_{k-1}\times d_k\times r_k} \,:\, \|C\|_{\ell^0_{\boldsymbol{\beta}_Q}} \le r \text{ with } \boldsymbol{\beta}_Q^2 := \operatorname{diag}(Q^\intercal \operatorname{diag}(\boldsymbol{\omega}^2)Q)\}.$$

Note that the new definition of $\mathcal{C}_{Q,r,\boldsymbol{\omega}}$ is a generalisation of the old definition to cases where the columns of $Q$ are not standard basis vectors. It may seem natural to assume that the memory footprint of the approximation grows from $(M-1)R + r$ for $v \in \mathcal{M}_{R,r,\boldsymbol{\omega}}$ to $MR^2d$ for $v \in \widetilde{\mathcal{M}}_{R,r,\boldsymbol{\omega}}$. However, since $\mathcal{M}_{R,r,\boldsymbol{\omega}} \subseteq \widetilde{\mathcal{M}}_{R,r,\boldsymbol{\omega}}$, an approximation of $v \in \mathcal{M}_{R,r,\boldsymbol{\omega}}$ may be possible in $\widetilde{\mathcal{M}}_{\tilde{R},r,\boldsymbol{\omega}}$ with $\tilde{R} \ll R$.

The representation of $v \in \widetilde{\mathcal{M}}_{R,r,\boldsymbol{\omega}}$ corresponds to the choice of

- a basis for the core space $Q$ as well as
- a weight sequence $\boldsymbol{\beta}_Q$.

In Theorem 5.12 we show that this choice ensures that a tensor with an $\ell^0_{\boldsymbol{\beta}_Q}$-sparse core is close to a sparse vector in $\ell^0_{\boldsymbol{\omega}}$. This is quite surprising since for any sparse core $C$ there exists an easy to construct $C$-dependent orthogonal basis $U$ such that all coefficients of the full tensor $UC$ are equal to $\|C\|_2$. This means that $UC$ is the least sparse tensor possible. However, using information about the weight sequence $\boldsymbol{\omega}$, we can construct a basis $Q$ and a weight sequence $\boldsymbol{\beta}_Q$ such that the full tensor $QC$ retains some of the sparsity of the core $C$.

**Remark 5.4.** Since $\mathcal{M}_{R,r,\boldsymbol{\omega}} \subseteq \widetilde{\mathcal{M}}_{R,r,\boldsymbol{\omega}}$, the approximation error for $\widetilde{\mathcal{M}}_{R,r,\boldsymbol{\omega}}$ can be bounded from above by Corollary 4.10. To obtain a tighter bound, the total approximation error can be split into the low-rank approximation error and a subsequent weighted best $n$-term approximation of the core tensor

$$\|v - v_{\text{low-rank \& sparse}}\|_{L^2} \le \|v - v_{\text{low-rank}}\|_{L^2} + \|v_{\text{low-rank}} - v_{\text{low-rank \& sparse}}\|_{L^2}.$$

The first term is a classical low-rank approximation error, which is studied in [9, 10, 49] for $v \in H^k([0,1]^m)$ and in [9, 30] for $v \in H^{k_1}([0,1]^{d_1}) \otimes \cdots \otimes H^{k_m}([0,1]^{d_m})$. The second term is a sparse approximation error, which can in principle be bounded by applying the weighted Stechkin's lemma to the core tensor $C$ of $v_{\text{low-rank}} = QC$. This gives the bound

$$\|v_{\text{low-rank}} - v_{\text{low-rank \& sparse}}\|_{L^2} = \|(I - P_{J_n})C\|_{\ell^2} \le c(\boldsymbol{\beta}, q, n)^{-1}\|C\|_{\ell^q_{\boldsymbol{\beta}}} \le \bar{c}(\boldsymbol{\beta}, q, n)^{-1}\|C\|_{\ell^2_{\bar{\boldsymbol{\beta}}}}$$

for some $q < 2$ and weight sequences $\boldsymbol{\beta}$ and $\bar{\boldsymbol{\beta}}$. However, bounding $\|C\|_{\ell^2_{\bar{\boldsymbol{\beta}}}}$ in terms of some norm of the full tensor $\|QC\|_{\ell^2_{\bar{\boldsymbol{\omega}}}}$, for some arbitrary $\bar{\boldsymbol{\omega}}$, requires knowledge of the operator norm $\|Q^\intercal\|_{\ell^2_{\bar{\boldsymbol{\beta}}}\to\ell^2_{\bar{\boldsymbol{\omega}}}}$, which is unknown a priori. However, if $Q$ is $\bar{\boldsymbol{\omega}}$-orthogonal and $\bar{\boldsymbol{\beta}} = Q^\intercal\bar{\boldsymbol{\omega}}$, then the low-rank approximation can be carried out with respect to the stronger $\ell^2_{\bar{\boldsymbol{\omega}}}$-norm and we can bound

$$\|C\|^2_{\ell^2_{\bar{\boldsymbol{\beta}}}} = C^\intercal \operatorname{diag}(\bar{\boldsymbol{\beta}}^2)C = (QC)^\intercal \operatorname{diag}(\bar{\boldsymbol{\omega}}^2)QC = \|v_{\text{low-rank}}\|^2_{\ell^2_{\bar{\boldsymbol{\omega}}}} \le \|v\|^2_{\ell^2_{\bar{\boldsymbol{\omega}}}}.$$

But this requires the low-rank approximation to be carried out with respect to a stronger norm than $L^2$. To the knowledge of the authors no rates for this are known.

**Remark 5.5** (Rank bounds for mixed Sobolev spaces)**.** Consider a function $u$ of $M$ variables and the corresponding sequence of coefficients $\boldsymbol{u} \in \mathbb{R}^{\mathbb{N}} \otimes \cdots \otimes \mathbb{R}^{\mathbb{N}}$ with respect to a tensor product basis. Moreover, suppose that $\boldsymbol{u}$ is $\ell^2_{\bar{\boldsymbol{\omega}}}$-summable with respect to the product weight sequence $\bar{\boldsymbol{\omega}} := \boldsymbol{\omega}^{\otimes M}$ with $\boldsymbol{\omega}_j := (j+1)^k$. The space $\ell^2_{\bar{\boldsymbol{\omega}}}$ captures the regularity of the mixed Sobolev spaces $H^{k,\mathrm{mix}} \simeq H^k \otimes \cdots \otimes H^k$. To bound the rank of $u$ by means of weighted sparsity we utilise the best $n$-term approximation rates for $\boldsymbol{u}$ from Remark 2.13. Recall that $r$-term approximation in a product basis can be represented with rank $r$ in the CP format and that the rank of any tree-based format is upper bounded by the CP rank (but may indeed be much smaller). This naïve bound yields (up to logarithmic factors) the best rank-$r$ approximation rates

$$\|u - u_r\|_{H^j} \leq r^{-(k-j)} \|u\|_{H^k}$$

for all $0 \leq j < k$. For $j = 0$, these bounds slightly extend the rates from [27, 28] but are worse than the more recent $r^{-2k}$ rates that are derived in [30] for the rank in the tensor train format. We conjecture that sparsity implies simple rank bounds with sparse components and a subsequent rounding can reduce the rank from $r^2$ to $r$. Note, however, that both rank bounds imply roughly the same approximation rates. By Theorem 4.9, the number of parameters that are needed to represent the best $r$-term approximation in the sparse tensor train format is bounded by $N = Mr$. This implies an approximation rate of $(\frac{N}{M})^{-k}$. If we consider the best rank-$r$ rate $r^{-2k}$ from [30], and assume that the component tensors in the corresponding tensor train representation are dense, then the number of parameters scales like $N \in \mathcal{O}(Mr^2)$ and we obtain the same approximation rate of $(\frac{N}{M})^{-k}$. We also remark that the best $r$-term approximation rates crucially depend on the chosen basis, while the ranks do not. Therefore, the rank-$r$ approximation rates that are obtained by this method can only provide upper bounds. Moreover, the ordering of the modes matters for the ranks of a tensor train representation. This is not reflected in these simple bounds, where the ordering is only important for an anisotropic choice of weight sequences.

**Remark 5.6** (Constructive rank bounds for Sobolev spaces)**.** Similar to the hierarchical SVD (or PCA) that can be used to construct classical low-rank representation, we can perform the weighted LASSO hierarchically to construct simultaneously sparse and low-rank representations. In this remark we demonstrate this procedure for the Tucker decomposition. Consider a function $u$ of $M$ variables and the corresponding sequence of coefficients $\boldsymbol{u} \in \mathbb{R}^{\mathbb{N}} \otimes \cdots \otimes \mathbb{R}^{\mathbb{N}}$ with respect to a tensor product basis. Moreover, suppose that $\boldsymbol{u}$ is $\ell^2_{\bar{\boldsymbol{\omega}}}$-summable with respect to the weight sequence $\bar{\boldsymbol{\omega}} := \sum_{m=1}^{M} \mathbf{1}^{\otimes(m-1)} \otimes \boldsymbol{\omega} \otimes \mathbf{1}^{\otimes(M-m)}$ with $\boldsymbol{\omega}_j := (j+1)^k$. The space $\ell^2_{\bar{\boldsymbol{\omega}}}$ captures the regularity of the standard Sobolev spaces

$$H^k \simeq \bigcap_{m=1}^{M} (L^2)^{\otimes(m-1)} \otimes H^k \otimes (L^2)^{\otimes(M-m)}.$$

We can bound the Tucker-rank of $u$ by means of weighted sparsity. The method that we use to derive our rank bounds is constructive and proceeds analogously to the HOSVD algorithm. We define for every $m = 1, \ldots, M$ the matricisation

$$(\boldsymbol{u}^{(m)}_j)_{i_1,\ldots,i_{M-1}} = \boldsymbol{u}_{i_1,\ldots,i_{m-1},j,i_m,\ldots,i_{M-1}},$$

which we interpret as a sequence of tensors of order $M-1$. Applying the weighted Stechkin lemma to the sequence $\boldsymbol{u}^{(m)}$, we select $r$ many $(M-1)$-dimensional "slices" of $\boldsymbol{u}$ and set the remaining slices to zero. This results in a new tensor which we denote by $\widetilde{\boldsymbol{u}}^{(m)}$. This construction can be performed sequentially for every $m = 1, \ldots, M$, leading to the sequence of approximations

$$\boldsymbol{u} =: \widetilde{\boldsymbol{u}}^{(0)} \rightsquigarrow \widetilde{\boldsymbol{u}}^{(1)} \rightsquigarrow \ldots \rightsquigarrow \widetilde{\boldsymbol{u}}^{(M)}.$$

The approximation error of this scheme is given by the telescoping sum

$$\begin{aligned}\|\boldsymbol{u} - \widetilde{\boldsymbol{u}}^{(M)}\|_{\ell^2}^2 &= \sum_{m=1}^{M} \|\widetilde{\boldsymbol{u}}^{(m-1)} - \widetilde{\boldsymbol{u}}^{(m)}\|_{\ell^2}^2 \\ &\lesssim \sum_{m=1}^{M} r^{-2k} \|(\mathbf{1}^{\otimes(m-1)} \otimes \boldsymbol{\omega} \otimes \mathbf{1}^{\otimes(M-m)})\widetilde{\boldsymbol{u}}^{(m-1)}\|_{\ell^2}^2,\end{aligned}$$

where the inequality follows from the weighted Stechkin lemma applied to the tensor-valued sequence $\widetilde{\boldsymbol{u}}^{(m-1)}$, which is weighted by $\boldsymbol{\omega}$ and a subsequent application of Lemma 2.11. Bounding $\|(\mathbf{1}^{\otimes(k-1)} \otimes \boldsymbol{\omega} \otimes \mathbf{1}^{\otimes M-k-1})\widetilde{\boldsymbol{u}}^{(k)}\|_{\ell^2} \leq \|\boldsymbol{u}\|_{\ell^2_{\boldsymbol{\omega}}}$ yields the simplified expression

$$\|u - u_r\|_{L^2} = \|\boldsymbol{u} - \widetilde{\boldsymbol{u}}^{(M)}\|_{\ell^2} \lesssim \sqrt{M} r^{-k} \|\boldsymbol{u}\|_{\ell^2_{\boldsymbol{\omega}}}. = \sqrt{M} r^{-k} \|u\|_{H^k}.$$

But in contrast to the model class in Section 4 the matrices $Q \in \widetilde{\mathcal{Q}}_{k,R,\boldsymbol{\omega}}$ are not spanned by a subbasis of the standard product basis. As a consequence, $\widetilde{\mathcal{M}}_{R,r,\boldsymbol{\omega}}$ is no longer a subset of the sparse vectors $\ell^0_{\boldsymbol{\omega}}$ and Theorem 1.5 can no longer be applied directly, as it was done in the proof of Corollary 4.12. Instead, we rely on the following strong property: If the RIP is satisfied at a point, it is satisfied in a small neighbourhood of that point. This is stated formally in the subsequent lemma.

**Lemma 5.7.** *Let $\delta, \tau \in [0,1)$. Then there exists a constant $\varepsilon \geq 8\tau$ such that for every $a \in L^\infty_w$*

$$\mathrm{RIP}_{\{a\}}(\delta) \Rightarrow \mathrm{RIP}_{B_{L^\infty_w}(a, \|a\|\tau)}(\delta + \varepsilon),$$

*with $\varepsilon \leq 15\tau$, if $\delta \leq \frac{1}{2}$ and $\tau \leq \frac{1}{4}$. For $\delta \leq \frac{1}{2}$ and $\tau \leq \frac{\delta}{15}$, this implies*

$$\mathrm{RIP}_{\{a\}}(\delta) \Rightarrow \mathrm{RIP}_{B_{L^\infty_w}(a, \|a\|\tau)}(2\delta).$$

To prove this lemma, we require the following result.

**Lemma 5.8.** *Let $a, b \in \mathcal{V}$ be bounded with respect to $\|\bullet\|_{w,\infty}$ and define $\widetilde{a} := \frac{a}{\|a\|}$ and $\widetilde{b} := \frac{b}{\|b\|}$. Moreover, let $\|\bullet\|_*$ denote either the norm $\|\bullet\|$ or the empirical norm $\|\bullet\|_n$. Then*

$$\|\widetilde{a} - \widetilde{b}\|_* \leq \frac{1 + \|\widetilde{b}\|_*}{\|a\|} \|a - b\|_{L^\infty_w}.$$

**Proof.** By triangle and reverse triangle inequality, it holds that

$$\begin{aligned}\left\| \frac{a}{\|a\|} - \frac{b}{\|b\|} \right\|_* \leq \frac{1}{\|a\|} \left( \|a - b\|_* + \|b - \tfrac{\|a\|}{\|b\|} b\|_* \right) &= \frac{\|a - b\|_*}{\|a\|} + \frac{|\|b\| - \|a\||}{\|a\|} \frac{\|b\|_*}{\|b\|} \\ &\leq \frac{\|a - b\|_*}{\|a\|} + \frac{\|a - b\|}{\|a\|} \frac{\|b\|_*}{\|b\|}.\end{aligned}$$

The claim follows, since both $\|\bullet\|$ and $\|\bullet\|_n$ are dominated by $\|\bullet\|_{L^\infty_w}$. ∎

**Proof.** [Proof of Lemma 5.7] Let $B := B_{\|\bullet\|_{w,\infty}}(a, r)$ with $r := \|a\|\tau$ and define $\widetilde{b} := \frac{b}{\|b\|}$ for any $b \in B$. We want to show that

$$\mathrm{RIP}_B(\delta + \varepsilon) \Leftrightarrow |\|\widetilde{b}\|^2 - \|\widetilde{b}\|_n^2| \leq \delta + \varepsilon \text{ for all } b \in B, \tag{5.1}$$

given that $\mathrm{RIP}_{\{a\}}(\delta) \Leftrightarrow |\|\widetilde{a}\|^2 - \|\widetilde{a}\|_n^2| \leq \delta$ holds. For this, let $\|\bullet\|_*$ denote either $\|\bullet\|$ or $\|\bullet\|_n$ and observe that for any $b \in B$ it holds that

$$\left| \|\widetilde{a}\|_*^2 - \|\widetilde{b}\|_*^2 \right| \leq (\|\widetilde{a}\|_* + \|\widetilde{b}\|_*) \|\widetilde{a} - \widetilde{b}\|_* \leq (\|\widetilde{a}\|_* + \|\widetilde{b}\|_*)(1 + \|\widetilde{b}\|_*)\tau, \tag{5.2}$$

where the last inequality follows from Lemma 5.8 and the assumption $\|a - b\|_{L^\infty_w} \leq r = \|a\|\tau$. By assumption, it holds that $\|\widetilde{a}\|_n \leq \sqrt{1 + \delta}$ and using the fact that $\|a - b\|_* \leq \|a - b\|_{L^\infty_w} \leq r$, we can

bound

$$\|\widetilde{b}\|_n = \frac{\|b\|_n}{\|b\|} \leq \frac{\|a\|_n + r}{\|a\| - r} \leq \frac{\sqrt{1+\delta}\|a\| + r}{\|a\| - r} = \frac{\sqrt{1+\delta} + \tau}{1-\tau}.$$

Inserting these estimates into equation (5.2) gives the bounds

$$\left|\|\widetilde{a}\|^2 - \|\widetilde{b}\|^2\right| \leq 4\tau \qquad \text{and} \qquad \left|\|\widetilde{a}\|_n^2 - \|\widetilde{b}\|_n^2\right| \leq 4\tau c,$$

with $c := \frac{1+\delta}{(1-\tau)^2}$. We can now use the triangle inequality to prove (5.1) via

$$\begin{aligned}
|\|\widetilde{b}\|^2 - \|\widetilde{b}\|_n^2| &\leq |\|\widetilde{b}\|^2 - \|\widetilde{b}\|_n^2 - (\|\widetilde{a}\|^2 - \|\widetilde{a}\|_n^2)| + |\|\widetilde{a}\|^2 - \|\widetilde{a}\|_n^2| \\
&\leq |\|\widetilde{b}\|^2 - \|\widetilde{a}\|^2| + |\|\widetilde{b}\|_n^2 - \|\widetilde{a}\|_n^2| + |\|\widetilde{a}\|^2 - \|\widetilde{a}\|_n^2| \\
&\leq \underbrace{4\tau + 4\tau c}_{=:\varepsilon} + \delta.
\end{aligned}$$

To obtain the lower bound for $\varepsilon$, observe that $c \geq 1$ with equality when $\delta = \tau = 0$. Therefore, $\varepsilon = 4(1+c)\tau \geq 8\tau$. To obtain the other bound, observe that the function $(\delta, \tau) \mapsto c(\delta, \tau)$ is increasing in both arguments. Therefore $c(\delta, \tau) \leq c(\frac{1}{2}, \frac{1}{4}) = \frac{8}{3}$ for any $\delta \leq \frac{1}{2}$ and $\tau \leq \frac{1}{4}$. This results in the loose upper bound

$$\varepsilon \leq 15\tau.$$

The special case $\tau \leq \frac{\delta}{15}$ follows immediately. ∎

For the sake of brevity, let $A := \mathcal{M}_{R,r,\boldsymbol{\omega}}$ and $B := B_{\ell^0_{\boldsymbol{\omega}}}(0, r)$. In the preceding section, we used the fact that $A \subseteq B$ to trivially obtain the restricted isometry property of $A$ from that of $B$. Lemma 5.7 implies that this inclusion is not necessary for the RIP of $B$ to extend to $A$ if the set $A$ is close enough to $B$. To make this intuition rigorous, we define the scale-invariant non-symmetric distance function

$$d_{\mathrm{si}L^\infty_w}(A, B) := \sup_{a \in A} d_{L^\infty_w}(\mathrm{Cone}(a), U(B)),$$

where the distance function $d_{L^\infty_w}(A, B)$ is defined as

$$d_{L^\infty_w}(A, B) := \inf_{a \in A} \inf_{b \in B} \|a - b\|_{L^\infty_w}.$$

With this definition, we can formulate the following theorem.

**Theorem 5.9.** *Let $\delta, r \in [0, 1)$ and assume that $d_{\mathrm{si}L^\infty_w}(A, B) \leq r$. Then there exists $\varepsilon \geq 8r$ such that*

$$\mathrm{RIP}_B(\delta) \Rightarrow \mathrm{RIP}_A(\delta + \varepsilon),$$

*with $\varepsilon \leq 15r$ if $\delta \leq \frac{1}{2}$ and $r \leq \frac{1}{4}$. For $\delta \leq \frac{1}{2}$ and $r \leq \frac{\delta}{15}$, this implies*

$$\mathrm{RIP}_B(\delta) \Rightarrow \mathrm{RIP}_A(2\delta).$$

**Proof.** Since $d_{\mathrm{si}L^\infty_w}(A, B) \leq r$, it holds that for all $a \in A$ there exists $t \in (0, \infty)$ and $b \in U(B)$ such that $d_{L^\infty_w}(ta, b) \leq r$. Assuming $\mathrm{RIP}_B(\delta)$, Lemma 5.7 guarantees that there exists a constant $\varepsilon$ satisfying the given bounds such that

$$\mathrm{RIP}_B(\delta) \Rightarrow \mathrm{RIP}_{U(B)}(\delta) \Rightarrow \mathrm{RIP}_{\{b\}}(\delta) \Rightarrow \mathrm{RIP}_{B_{L^\infty_w}(b,r)}(\delta + \varepsilon) \Rightarrow \mathrm{RIP}_{\{ta\}}(\delta + \varepsilon) \Rightarrow \mathrm{RIP}_{\{a\}}(\delta + \varepsilon).$$

This implies $\mathrm{RIP}_{\{a\}}(\delta + \varepsilon)$ for all $a \in A$ and consequently $\mathrm{RIP}_A(\delta + \varepsilon)$. ∎

**Remark 5.10.** Note that

$$d_{\mathrm{si}L^\infty_w}(A, B) := \sup_{a \in A} d_{L^\infty_w}(\mathrm{Cone}(a), U(B)) \leq \sup_{\widetilde{a} \in U(A)} d_{L^\infty_w}(\widetilde{a}, U(B)) =: d_{\mathrm{h}}(U(A), U(B)),$$

where $d_{\mathrm{h}}$ is the directed Hausdorff distance between sets. This bound can be used in conjunction with Theorem 5.9 to provide a simple proof of one of the major corollaries in [54]. Consider the setting of

Proposition 1.4. Assume that $\mathcal{M}$ is a manifold with strictly positive reach $R := \operatorname{rch}(\mathcal{M}, u_{\mathcal{M}})$ at point $u_{\mathcal{M}}$, and define $\mathcal{M}_r := \mathcal{M} \cap B(u_{\mathcal{M}}, r)$ for any $r \leq R$. By Proposition 16 in [54], it holds that

$$d_{\mathrm{h}}(U(\{u_{\mathcal{M}}\} - \mathcal{M}_r), U(\mathbb{T}_{u_{\mathcal{M}}}\mathcal{M})) \leq \frac{r}{R}.$$

From this follows that there exists $\varepsilon > 0$ such that

$$\mathrm{RIP}_{\mathbb{T}_{u_{\mathcal{M}}}\mathcal{M}}(\delta) \Rightarrow \mathrm{RIP}_{\{u_{\mathcal{M}}\}-\mathcal{M}_r}(\delta + \varepsilon).$$

This means that if the RIP holds for the tangent space at $u_{\mathcal{M}}$ then it also holds for a neighbourhood of $u_{\mathcal{M}}$ in $\mathcal{M}$. This is precisely the property that is required in Proposition 1.4.

**Remark 5.11.** Note that Theorem 5.9 may also be used to verify $\mathrm{RIP}_A(2\delta)$ by checking $\mathrm{RIP}_B(\delta)$ for a finite subset $B \subseteq A$ with $d_{\mathrm{si}L^\infty_w}(A, B) \leq \frac{\delta}{15}$.

The preceding theorem can be utilised to prove the RIP for our model class of semi-sparse tensors $\widetilde{\mathcal{M}}_{R,r,\boldsymbol{\omega}}$. This is done in the subsequent theorem, which chooses $A := \widetilde{\mathcal{M}}_{R,r,\boldsymbol{\omega}}$ and $B := B_{\widetilde{r}}(\ell^0_{\boldsymbol{\omega}})$ and shows that

$$d_{\mathrm{si}L^\infty_w}(A, B) \leq \mathcal{O}\Big(\sqrt{\frac{r}{\widetilde{r}}}\Big).$$

**Theorem 5.12.** *Let $\{B_j\}_{j\in[D]}$ be orthonormal with respect to the measure $\rho$ and let $w \geq 0$ be any weight function satisfying $\|w^{-1}\|_{L^1} = 1$. Assume the weight sequence satisfies $\boldsymbol{\omega}_j \geq \|w^{1/2}B_j\|_{L^\infty}$ and fix $c, r > 0$ and $\widetilde{r} := (1 + c^2)\|\omega^{-1}\|^2_{\ell^{2/3}} r$. Then it holds that*

$$d_{\mathrm{si}L^\infty_w}(\widetilde{\mathcal{M}}_{R,r,\boldsymbol{\omega}}, B_{\ell^0_{\boldsymbol{\omega}}}(0, \widetilde{r})) \leq \frac{1}{c}.$$

*Hence, if $\delta \leq \frac{1}{2}$ and $c \geq \frac{15}{\delta}$, then $\mathrm{RIP}_{B_{\ell^0_{\boldsymbol{\omega}}}(0,\tilde{r})}(\delta)$ implies $\mathrm{RIP}_{\tilde{\mathcal{M}}_{R,r,\boldsymbol{\omega}}}(2\delta)$.*

**Proof.** Let $A \in \widetilde{\mathcal{M}}_{R,r,\boldsymbol{\omega}}$. Since $\boldsymbol{\omega}_\nu \geq \|B_\nu\|_{L^\infty}$, Corollary 2.8 states that for every $\widetilde{r} > 0$ there exists $\widetilde{J} \subseteq \mathbb{N}$ and $\widetilde{A} := P_{\tilde{J}}A$ such that $\widetilde{A} \in B_{\ell^0_{\boldsymbol{\omega}}}(0, \widetilde{r})$ and

$$\|A - \widetilde{A}\|_{L^2} \leq \|A - \widetilde{A}\|_{L^\infty} \leq \|A - \widetilde{A}\|_{\ell^1_{\boldsymbol{\omega}}} \leq \widetilde{r}^{-1/2}\|A\|_{\ell^{2/3}_{\boldsymbol{\omega}^2}}. \tag{5.3}$$

The final $\ell^{2/3}_{\boldsymbol{\omega}^2}$-norm can be bounded by Lemma 2.11 via

$$\|A\|_{\ell^{2/3}_{\boldsymbol{\omega}^2}} \leq \|\boldsymbol{\omega}^{-3}\|_{\ell^{2/3}_{\boldsymbol{\omega}^2}}\|A\|_{\ell^2_{\boldsymbol{\omega}^3}} = \|\boldsymbol{\omega}^{-1}\|_{\ell^{2/3}}\|A\|_{\ell^2_{\boldsymbol{\omega}^3}}. \tag{5.4}$$

Recall that $A \in \widetilde{\mathcal{M}}_{R,r,\boldsymbol{\omega}}$ can be written as $A = QC$ with $Q \in \widetilde{\mathcal{Q}}_{k,R,\boldsymbol{\omega}}$ for some $k \in [M]$ and $C \in \mathcal{C}_{Q,r,\boldsymbol{\omega}}$. Thus,

$$\|A\|^2_{\ell^2_{\boldsymbol{\omega}^3}} = (QC)^\intercal \operatorname{diag}(\boldsymbol{\omega}^6)(QC) = C^\intercal \operatorname{diag}(\beta_Q^2)C = \|C\|^2_{\ell^2_{\beta_Q}}. \tag{5.5}$$

Now let $J := \operatorname{supp}(C)$ and bound

$$\|C\|^2_{\ell^2_{\beta_Q}} = \sum_{j\in J} \beta^2_{Q,j}C_j^2 \leq \sum_{j\in J}\Big(\sum_{j\in J}\beta^2_{Q,j}\Big)C_j^2 = \|C\|_{\ell^0_{\beta_Q}}\|C\|^2_{\ell^2} \leq r\|C\|^2_{\ell^2} = r\|A\|^2_{\ell^2}. \tag{5.6}$$

Combining equations (5.3), (5.4), (5.5) and (5.6) results in the bound

$$\|A - \widetilde{A}\|_{L^2} \leq \|A - \widetilde{A}\|_{L^\infty} \leq \sqrt{c_1\tfrac{r}{\widetilde{r}}}\|A\|_{L^2},$$

with $c_1 := \|\boldsymbol{\omega}^{-1}\|^2_{\ell^{2/3}}$. Finally, recall that $\widetilde{A} := P_{\tilde{J}}A$ and hence

$$\|\widetilde{A}\|^2_{L^2} = \|A\|^2_{L^2} - \|A - \widetilde{A}\|^2_{L^2} \geq \tfrac{\widetilde{r}-c_1 r}{\widetilde{r}}\|A\|^2_{L^2}.$$

This means that for every $A \in \widetilde{M}_{R,r,\boldsymbol{\omega}}$ there exists $\widetilde{A} \in B_{\widetilde{r}}(\ell^0_{\boldsymbol{\omega}})$ such that

$$\|A - \widetilde{A}\|_{L^\infty} \leq \sqrt{c_1\tfrac{r}{\widetilde{r}}}\|A\|_{L^2} \leq \sqrt{c_1\tfrac{r}{\widetilde{r}}} \cdot \sqrt{\tfrac{\widetilde{r}}{\widetilde{r}-c_1 r}}\|\widetilde{A}\|_{L^2} = \sqrt{\tfrac{c_1 r}{\widetilde{r}-c_1 r}}\|\widetilde{A}\|_{L^2} = \tfrac{1}{c}\|\widetilde{A}\|_{L^2},$$

where the final equality follows from the choice $\widetilde{r} := (1+c^2)\|\boldsymbol{\omega}^{-1}\|_{\ell^{2/3}}^2 r = c^2 c_1 r + c_1 r$. This shows that for all $A \in \widetilde{\mathcal{M}}_{R,r,\boldsymbol{\omega}}$ there exists a constant $t := \|\widetilde{A}\|_{L^2}^{-1} > 0$ and an element $\bar{A} := t\widetilde{A} \in U(B_{\ell^0_{\boldsymbol{\omega}}}(0,\widetilde{r}))$ such that $d_{L^\infty}(tA, \bar{A}) \leq \frac{1}{c}$. In other words,

$$d_{\mathrm{si}L^\infty}(\widetilde{\mathcal{M}}_{R,r,\boldsymbol{\omega}}, B_{\ell^0_{\boldsymbol{\omega}}}(0,\widetilde{r})) \leq \frac{1}{c}.$$

The claim now follows from Theorem 5.9. ∎

Even though Theorem 5.12 is valid for any weight sequence $\boldsymbol{\omega}$, an increasing sequence $\boldsymbol{\omega}$ is necessary in practice. If $\boldsymbol{\omega} \propto 1$, then $\widetilde{r} = (1+c^2)\|\boldsymbol{\omega}^{-1}\|_{\ell^{2/3}}^2 r \propto (1+c^2)\dim(\boldsymbol{\omega})^3 r$, where the dimension of the weight vector $\boldsymbol{\omega}$ is the dimension of the ambient tensor product space. Hence, applying Theorem 5.12 with a constant weight sequence would require the RIP to hold for the entire ambient tensor product space.

The nestedness property of the model classes $\widetilde{\mathcal{M}}_{R,r,\boldsymbol{\omega}}$ can be proved in the same way as for the model class $\mathcal{M}_{R,r,\boldsymbol{\omega}}$.

**Proposition 5.13.** *It holds that* $\widetilde{\mathcal{M}}_{R,r,\boldsymbol{\omega}} - \widetilde{\mathcal{M}}_{R,r,\boldsymbol{\omega}} \subseteq \widetilde{\mathcal{M}}_{2R,2r,\boldsymbol{\omega}}$.

This allows the application of Proposition 1.4. As in the preceding section, we can use Theorem 1.5 to provide a bound for the required number of samples when the model class $\widetilde{\mathcal{M}}_{R,r,\boldsymbol{\omega}}$ is used in the optimisation problem (1.5). As before, let $b : Y \to \mathbb{R}^d$ be a vector of $L^2(Y,\rho)$-orthonormal basis functions, define the tensor product basis $B(y) := b(y_1) \otimes \cdots \otimes b(y_M)$ and suppose that the weight sequence $\boldsymbol{\omega}$ satisfies $\boldsymbol{\omega}_j \geq \|B_j\|_{L^\infty}$. Then the following proposition holds true.

**Corollary 5.14.** *Fix parameters $\gamma \in (0,1)$ and $\delta \in (0, \frac{1}{2})$. Let $\{B_j\}_{j\in[D]}$ be orthonormal with respect to the measure $\rho$ and let $w \geq 0$ be any weight function satisfying $\|w^{-1}\|_{L^1} = 1$. Assume the weight sequence satisfies $\boldsymbol{\omega}_j \geq \|w^{1/2}B_j\|_{L^\infty}$ and fix $\widetilde{r} := (1+c^2)\|\boldsymbol{\omega}^{-1}\|_{\ell^{2/3}}^2 r$ for some $c > \frac{15}{\delta}$. Then, if*

$$n \geq C\delta^{-2}\widetilde{r}\max\{\log^3(\widetilde{r})\log(d^M), -\log(\gamma)\}$$

*and $y_1, \ldots, y_n$ are drawn independently from $w^{-1}\rho$, the probability of* $\mathrm{RIP}_{\tilde{\mathcal{M}}_{R,r,\boldsymbol{\omega}}}(2\delta)$ *exceeds* $1-\gamma$.

**Proof.** Under the given assumptions, Theorem 1.5 guarantees $\mathrm{RIP}_{B_{\ell^0_{\boldsymbol{\omega}}([d]^M)}(0,\widetilde{r})}(\delta)$. This implies $\mathrm{RIP}_{\tilde{\mathcal{M}}_{R,r,\boldsymbol{\omega}}}(2\delta)$ by Theorem 5.12. ∎

Although it is not clear how to write an algorithm that remains in this model class, this is not a significant drawback, since we can again choose $r$ by cross-validation and $R$ by standard rank adaptation strategies.

### 5.2. **Numerical method**

We call the resulting algorithm semisparse ALS (SSALS). The only difference of this method to the sparse ALS (Algorithm 2) is the usage of the $\boldsymbol{\omega}$-orthogonal QC decomposition instead of a sparse QC decomposition. Due to this change, the SSALS looses the intrinsic rank-adaptivity of the SALS. But since SSALS is stable by design, the tensor train rank of the coefficient tensor can be chosen arbitrary. Note that from an approximation error point of view, it would even be optimal to perform SSALS on a full rank tensor, which is infeasible due to the size of the resulting component tensors. We hence propose to implement a rank-adaptive algorithm that is based on the rank-adaptation strategy proposed in [25]. This approach splits the sequence of singular values of a singular value decomposition into two groups. The first group contains all singular values that exceed a certain significance threshold and the second group contains all remaining singular values. By fixing the size of the second group,

dropping the smallest singular values or adding small random singular values if necessary, adaptivity is achieved. Moreover, since the second group is assumed insignificant, the corresponding singular vectors can be perturbed randomly without adversely affecting the approximation error. This allows to randomly explore the space of singular vectors in order to find those that are necessary to represent the sought function. If a singular vector in the second group is important to represent the sought function, the corresponding singular value increases during optimisation and is eventually assigned to the first group.

An important aspect of any iterative algorithm is its initialisation. Since both algorithms are adaptive by nature, we can choose as initialisation the constant function returning the empirical mean.

To monitor convergence, we split the data set into training and validation sets, computing the least squares residual on both sets. We terminate after a fixed number of iterations (50 in the experiments below) or when the training set error does not decrease for 3 consecutive iterations. (The cross-validation procedure in the microstep prevents overfitting.) During the iteration, we store the iterate with minimum validation set error and return it once the algorithm terminates.

More details can be found in our source code at `github.com/ptrunschke/sparse_als`.

## 6. **Experiments**

This section is concerned with numerical experiments that illustrate the practical performance of the sparse ALS algorithms derived from the theoretical results in the previous sections. We examine the reconstruction of a quantity of interest of the finite dimensional Darcy problem (1.1) with affine and log-affine coefficients. From Theorem 1.3 it is known that the solution lies in an exponentially weighted $\ell^2$ space. As a consequence, a weighted LASSO as used in the SALS should (at least theoretically) provide good approximation rates. Since the bases of the micro steps may become very large, as discussed in Section 5, we modify the SALS to terminate after a fixed maximal time.

The source code of the implementation is available at `github.com/ptrunschke/sparse_als`. Moreover, we compare our results to the highly optimised `tensap` library [42], which can be found at `github.com/anthony-nouy/tensap`.

### 6.1. **Affine Darcy equation**

Our first experiment is taken from [11], where a weighted $\ell^1$ minimisation was used. We consider model problem (1.1) on the unit interval $D = [0,1]$ and parameter domain $Y = [-1,1]^L$ with $L = 20$. We consider the forcing term $f \equiv 10$ and the diffusion coefficient

$$a(x,y) := \frac{1}{10} + \frac{\pi^2}{3} + \sum_{m=1}^{L} k^{-2} a_m(x) y_m,$$

where $a_{2m-1}(x) = \cos(m\pi x)$ and $a_{2m} = \sin(m\pi x)$. The PDE in its variational form is solved on a uniform grid with 50 nodes using conforming $P1$ finite elements. In this first experiment we consider the quantity of interest

$$U(y) := \int_D u(x,y)\,\mathrm{d}x.$$

We use the probability measure $\rho = \frac{1}{2^L}\,\mathrm{d}y$ and weight function $w \equiv 1$ (cf. (1.4)) and search for the best approximation with respect to $\|\bullet\|_{L^2(Y,\rho)}$, using a product basis with $d = 20$ Legendre polynomials in each variable. Concerning the weight sequence, we utilise the smallest possible choice $\boldsymbol{\omega}_\alpha := \|B_\alpha\|_{L^\infty}$. Note that this is not the exponential weighting, that we could have used according to Lemma 2.1 and Theorem 1.2. Numerical results for the proposed algorithms for the empirical best-approximation of $U$ is provided in Table 6.1.

| | $n = 50$ | $n = 100$ | $n = 500$ | $n = 1000$ |
|---|---|---|---|---|
| SALS (ours) | $[5.9 \cdot 10^{-4}, 4.0 \cdot 10^{-3}]$ | $[4.4 \cdot 10^{-4}, 3.5 \cdot 10^{-3}]$ | $[1.2 \cdot 10^{-3}, 2.9 \cdot 10^{-3}]$ | $[1.2 \cdot 10^{-3}, 2.8 \cdot 10^{-3}]$ |
| SSALS (ours) | $[2.3 \cdot 10^{-3}, 4.0 \cdot 10^{-3}]$ | $[2.3 \cdot 10^{-3}, 3.8 \cdot 10^{-3}]$ | $[2.0 \cdot 10^{-3}, 3.1 \cdot 10^{-3}]$ | $[1.3 \cdot 10^{-3}, 3.0 \cdot 10^{-3}]$ |
| tensap | $[8.6 \cdot 10^{-4}, 4.9 \cdot 10^{-3}]$ | $[4.9 \cdot 10^{-4}, 3.5 \cdot 10^{-3}]$ | $[1.0 \cdot 10^{-4}, 2.8 \cdot 10^{-3}]$ | $[4.9 \cdot 10^{-5}, 8.4 \cdot 10^{-4}]$ |

TABLE 6.1. Relative $L^2$-approximation error for the quantity of interest in Section 6.1. The relative error in the $L^2$-norm is estimated on a test set of 1 000 independent samples. The experiments are performed 10 times and the 5% and 95% quantiles are displayed. All algorithms use the same samples to compute the empirical approximation (in each column) and the errors are always computed on the same test set. SALS and SSALS are compared to the `TreeBasedTensorLearning` procedure with basis adaptation from `tensap`.

### 6.2. **Log-affine Darcy equation**

The second example considers the Darcy equation with log-affine coefficient with $D = [0,1]^2$ and $Y = \mathbb{R}^L$ where $L = 20$. We define $f \equiv 1$ and

$$a(x,y) := \exp\left(S_k \sum_{m=1}^{L} m^{-k} \sin(\pi \lfloor \tfrac{m}{2} \rfloor x_1) \sin(\pi \lceil \tfrac{m}{2} \rceil x_2) y_m\right),$$

where $k \in \{1,2\}$, $S_1 := 2.4$ and $S_2 := 1.9$. As before, we solve the resulting PDE in its variational form on a uniform grid with $50 \times 50$ nodes using conforming $P1$ elements. The examined quantity of interest now is the coefficient of the most important POD mode corresponding to 20 000 sample points. The POD is computed by performing a SVD on the matrix of solution snapshots. The resulting singular vectors constitute an (almost) orthogonal basis and the coefficeient for the basis function that is associated with the largest singular value is used as the QoI. We choose $\rho = \mathcal{N}(0, I)$ as a multivariate standard normal distribution and $w$ such that $w^{-1}\rho$ is a multivariate centred normal distribution with variance $2I$ or $4I$.

In this experiment we search for the best approximation with respect to $\|\bullet\|_{L^2(Y,\rho)}$, using a basis of $d = 20$ Hermite polynomials in each mode. Similar to Theorem 1.3, the used weight sequence $\omega_\alpha := \|\sqrt{w} B_\alpha\|_{L^\infty}$ exhibits an exponential scaling. The results are depicted in Table 6.2 for $k = 1$ and in Tables 6.3 and 6.4 for $k = 2$.

### 6.3. **Discussion**

The numerical results illustrate that the obtained accuracy of the newly proposed sparse ALS algorithms SALS and SSALS is comparable to the highly optimised algorithm implemented in `tensap`, which we consider as base line. To understand the constraints of our sparse approach, recall the sample complexity bound

$$n \geq C\delta^{-2} r \max\{\log^3(r)\log(D), -\log(p)\}$$

from Theorem 1.5 with $D$ denoting the dimension of the full tensor product space. Given a fixed sample size $n$, stability parameter $\delta$ and probability $p$, this provides a heuristic upper bound on the weighted sparsity $r$ that can be achieved with our adaptive algorithms. To make this concrete, let $\delta = \sqrt{C}$ and $p \geq \frac{1}{2}$. Then

$$r \leq \frac{n}{\log(D)} = \frac{n}{20\log(20)} \leq \frac{n}{50},$$

| | $n = 50$ | $n = 100$ | $n = 500$ | $n = 1000$ |
|---|---|---|---|---|
| SALS (ours) | $[3.3 \cdot 10^{-1}, 1.3 \cdot 10^{0}]$ | $[2.4 \cdot 10^{-1}, 1.3 \cdot 10^{0}]$ | $[1.5 \cdot 10^{-1}, 3.4 \cdot 10^{-1}]$ | $[1.0 \cdot 10^{-1}, 1.9 \cdot 10^{-1}]$ |
| SSALS (ours) | $[3.4 \cdot 10^{-1}, 1.3 \cdot 10^{0}]$ | $[2.3 \cdot 10^{-1}, 8.8 \cdot 10^{-1}]$ | $[1.5 \cdot 10^{-1}, 4.0 \cdot 10^{-1}]$ | $[1.1 \cdot 10^{-1}, 3.2 \cdot 10^{-1}]$ |
| tensap | $[3.9 \cdot 10^{-1}, 8.3 \cdot 10^{-1}]$ | $[1.9 \cdot 10^{-1}, 8.8 \cdot 10^{-1}]$ | $[1.2 \cdot 10^{-1}, 4.5 \cdot 10^{-1}]$ | $[9.1 \cdot 10^{-2}, 3.2 \cdot 10^{-1}]$ |

Table 6.2. Relative $L^2$-approximation error for the quantity of interest for $k = 1$ in Section 6.2. The relative error in the $L^2$-norm is estimated on a test set of 1 000 independent samples. The experiments are performed 10 times and the 5% and 95% quantiles are displayed. All algorithms use the same samples to compute the empirical approximation (in each column) and the errors are always computed on the same test set. SALS and SSALS are compared to the `TreeBasedTensorLearning` procedure with basis adaptation from `tensap`.

| | $n = 50$ | $n = 100$ | $n = 500$ | $n = 1000$ |
|---|---|---|---|---|
| SALS (ours) | $[3.3 \cdot 10^{-2}, 2.3 \cdot 10^{-1}]$ | $[1.7 \cdot 10^{-2}, 8.3 \cdot 10^{-2}]$ | $[1.0 \cdot 10^{-2}, 3.4 \cdot 10^{-2}]$ | $[7.3 \cdot 10^{-3}, 2.8 \cdot 10^{-2}]$ |
| SSALS (ours) | $[3.3 \cdot 10^{-2}, 9.2 \cdot 10^{-2}]$ | $[2.1 \cdot 10^{-2}, 8.4 \cdot 10^{-2}]$ | $[8.4 \cdot 10^{-3}, 3.2 \cdot 10^{-2}]$ | $[8.3 \cdot 10^{-3}, 2.2 \cdot 10^{-2}]$ |
| tensap | $[3.8 \cdot 10^{-2}, 2.3 \cdot 10^{-1}]$ | $[2.1 \cdot 10^{-2}, 1.1 \cdot 10^{-1}]$ | $[1.1 \cdot 10^{-2}, 3.0 \cdot 10^{-2}]$ | $[6.4 \cdot 10^{-3}, 4.1 \cdot 10^{-2}]$ |

Table 6.3. Relative $L^2$-approximation error for the quantity of interest for $k = 2$ and a sampling distribution $w^{-1}\rho = \mathcal{N}(0, 2I)$ in Section 6.2. The relative error in the $L^2$-norm is estimated on a test set of 1 000 independent samples. The experiments are performed 10 times and the 5% and 95% quantiles are displayed. All algorithms use the same samples to compute the empirical approximation (in each column) and the errors are always computed on the same test set. SALS and SSALS are compared to the `TreeBasedTensorLearning` procedure with basis adaptation from `tensap`.

indicating that our adaptive algorithms are restricted to model classes with $r \leq \frac{n}{50}$. We suspect that this bound is implicitly imposed by the cross-validation inside the microstep. Since $1 \leq r$, this argument provides a theoretical explanation for the poor performance in the small-data regime. For $n = 50$, the bound $r \leq 1$ allows only to recover the mean since $B_0 \equiv 1$ is the only basis function with $\|B_j\|_{L^\infty} \leq 1$. This indicates that the problem in Subsection 6.1 is almost trivial to solve. In general, it can be seen from the numerical experiments that the errors decrease when the sample size increases. This is to be expected, since the probability of the restricted isometry property increases with the number of samples.

We should also note that, although the experiments seem to work very well, the automatic rank-adaptivity of SALS may fail in pathological scenarios. It is probably easy to construct an example where the automatic rank adaptation of the sparse QR can not increase the rank in a productive way. In this case, the semi-sparse ALS should nevertheless succeed, since the rank is adapted randomly.

In comparison to the baseline `tensap`, the shown results are quantitatively similar. We consider this very promising since the proposed algorithms only require simple modifications of the standard ALS

| | $n = 50$ | $n = 100$ | $n = 500$ | $n = 1000$ |
|---|---|---|---|---|
| SALS (ours) | $[3.6 \cdot 10^{-2}, 2.1 \cdot 10^{-1}]$ | $[7.2 \cdot 10^{-3}, 7.8 \cdot 10^{-2}]$ | $[2.4 \cdot 10^{-3}, 3.6 \cdot 10^{-2}]$ | $[5.1 \cdot 10^{-3}, 3.9 \cdot 10^{-2}]$ |
| SSALS (ours) | $[2.9 \cdot 10^{-2}, 1.9 \cdot 10^{-1}]$ | $[2.6 \cdot 10^{-2}, 1.3 \cdot 10^{-1}]$ | $[2.5 \cdot 10^{-2}, 1.8 \cdot 10^{-1}]$ | $[3.6 \cdot 10^{-2}, 2.2 \cdot 10^{-1}]$ |
| tensap | $[1.5 \cdot 10^{-2}, 2.2 \cdot 10^{0}]$ | $[4.1 \cdot 10^{-2}, 8.2 \cdot 10^{-1}]$ | $[4.4 \cdot 10^{-2}, 1.5 \cdot 10^{-1}]$ | $[2.2 \cdot 10^{-2}, 1.3 \cdot 10^{-1}]$ |

Table 6.4. Relative $L^2$-approximation error for the quantity of interest for $k = 2$ and a sampling distribution $w^{-1}\rho = \mathcal{N}(0, 4I)$ in Section 6.2. The relative error in the $L^2$-norm is estimated on a test set of 1 000 independent samples. The experiments are performed 10 times and the 5% and 95% quantiles are displayed. All algorithms use the same samples to compute the empirical approximation (in each column) and the errors are always computed on the same test set. SALS and SSALS are compared to the `TreeBasedTensorLearning` procedure with basis adaptation from `tensap`.

method. Note that the `tensap` algorithm adapts the basis functions with strategies based on leave-one-out cross validation [13, 14] or slope heuristics [40], and adapts the ranks based on a strategy similar to [25].

In general, the regularity of the considered function is encoded in $\boldsymbol{\omega}$ in the weighted Stechkin lemma. Since the target functionals in all experiments are (anisotropically) holomorphic functions, the best $n$-term sets $J_n$ are (anisotropic) balls in the index space. These are downward-closed sets matching exactly the basis selection strategy of `tensap`. It hence should almost be impossible to improve practical results on the selected model problems.

## 7. Conclusion

Motivated by the exceptional advantages that weighted sparsity brings to least squares approximation in terms of stability (cf. [46]), this paper considers the integration of weighted sparsity into tensor networks algorithms. To this end, we present a comprehensive framework for the weighted sparse and low-rank approximation of high-dimensional functions. Central to this approach is the introduction of a weighted version of Stechkin's lemma, which not only refines the classical convergence estimates but also better captures the decay properties inherent to many function classes, such as those arising in parametric PDEs. We harness the structure of the weighted Stechkin estimates using two different model classes and derive approximation rates. These theoretical insights are then translated into concrete numerical methods. Numerical experiments demonstrate that the proposed methods consistently yield near-optimal recovery of the target functions while mitigating the curse of dimensionality.

However, integrating weighted sparsity into tensor networks and utilising the results for parametric PDEs turns out to be too exhaustive for a single research paper, and many open questions remain.

In this work, we only derive weighted sparsity estimates for solutions of parametric PDEs with a single parameter. While we expect that these arguments will extend naturally to multi-parameter problems, this remains an open problem.

We present the proposed approximation algorithms without an analysis of numerical stability or a quantitative convergence analysis.

Translating the presented sparsity results to tight rank bounds for hierarchical and more general tensor formats also seems to be an exciting direction for future research. While Remark 5.5 provides the currently best rank bounds for general tensor network approximation in Sobolev spaces with mixed smoothness, it only yields suboptimal bounds for the special case of $L^2$-approximation with hierarchical

tensor formats (cf. [30]). Since the tensor networks in this construction exhibit sparse components and the ones in [30] do not, we conjecture that a subsequent rounding can reduce the rank from $r^2$ to $r$. However, a proof of this fact and an explicit construction remains an open problem. In Remark 5.6, we present an explicit and optimal (cf. [29]) construction for Tucker format approximation in Sobolev spaces on product domains.

Finally, we derive a novel result demonstrating that the restricted isometry property transfers automatically from any model class to any other "sufficiently close" model class (see Theorem 5.9). We believe that this insight provides a theoretical bridge for extending generalisation guarantees to more complex, less structured model classes, possibly even subsets of neural networks.

## Acknowledgements

Our code makes extensive use of the Python packages `numpy` [32], `scipy` [55], and `matplotlib` [35].

## Appendix A. Basic harmonic summation formulas

**Lemma A.1.** *For any $s > 0$ and $n \in \mathbb{N}$, we have $\frac{n^{s+1}}{s+1} \leq \sum_{k=1}^{n} k^s \leq \frac{(n+1)^{s+1}}{s+1}$.*

**Proof.** Both estimates rely on estimating the sum by an integral of the function $f(x) := x^s$. Since this function is convex, we know that the trapezoidal rule overestimates the integral and

$$\sum_{k=1}^{n} k^s = \frac{1}{2}(f(1) + f(n)) + \sum_{k=1}^{n-1} \frac{f(k) + f(k+1)}{2} \geq \frac{n^s + 1}{2} + \int_1^n f(x)\,\mathrm{d}x = \frac{n^s + 1}{2} + \frac{n^{s+1} - 1}{s+1}.$$

This gives the lower bound. For the upper bound, observe that $f$ is increasing and therefore a left Riemann sum underestimates the integral. Consequently,

$$\sum_{k=1}^{n} k^s \leq \int_1^{n+1} f(x)\,\mathrm{d}x = \frac{(n+1)^{s+1} - 1}{s+1} \leq \frac{(n+1)^{s+1}}{s+1}.$$

■

**Lemma A.2.** *For any $s > 1$ and $s \in \mathbb{N}$, we have $\frac{1}{s-1} k^{1-s} \leq \sum_{j=k}^{\infty} j^{-s} \leq \frac{s}{s-1} k^{1-s}$.*

**Proof.** Both estimates rely on estimating the sum by an integral of the function $f(x) := x^{-s}$. Since the function is decreasing, we know that a left Riemann sum overestimates the integral and

$$\sum_{j=k}^{\infty} j^{-s} \geq \int_k^{\infty} f(x)\,\mathrm{d}x = \frac{1}{s-1} k^{1-s}.$$

For the same reason, a right Riemann sum underestimates the integral and

$$\sum_{j=k}^{\infty} j^{-s} \leq f(k) + \int_k^{\infty} f(x)\,\mathrm{d}x = k^{-s} + \frac{k^{1-s}}{s-1} \leq \left(1 + \frac{1}{s-1}\right) k^{1-s} = \frac{s}{s-1} k^{1-s}.$$

■

## Appendix B. Example of a hierarchical basis

For $k = 1$ and $p = 2$ the equivalence of Example 2.3 is easy to see. In this case, we can use the basis

$$\phi_{\ell,j}(x) \propto \max\{1 - |2^\ell x - j|, 0\}, \qquad \|\phi_{\ell,j}\|_{L^2} = 1,$$

for any $\ell \in \mathbb{N}$ and odd $0 < j < 2^\ell$. If we represent $v \in W^{1,2}$ by successive $L^2$-projections onto the spaces $V_\ell := \operatorname{span}\{\phi_{\ell,j} : 0 < j < 2^\ell \text{ odd}\}$, it is easy to see that

$$\|v\|_{L^2}^2 = \left\| \sum_{\ell\in\mathbb{N}} \sum_j \boldsymbol{v}_{\ell,j}\phi_{\ell,j} \right\|_{L^2(\lambda)}^2 = \sum_{\ell\in\mathbb{N}} \left\| \sum_j \boldsymbol{v}_{\ell,j}\phi_{\ell,j} \right\|_{L^2(\lambda)}^2 = \sum_{\ell\in\mathbb{N}} \sum_j |\boldsymbol{v}_{\ell,j}|^2,$$

where the second equality follows due to the hierarchical projection and the third follows since $\phi_{\ell,j}$ have disjoint support for fixed $\ell$. Moreover, it is easy to see that all $\phi_{j,\ell}$ are orthogonal with respect to the $H_0^1$ semi inner product. This implies that

$$\|v\|_{H_0^1(\lambda)}^2 = \sum_{\ell\in\mathbb{N}} \sum_j |\boldsymbol{v}_{\ell,j}|^2 \|\phi_{\ell j}\|_{H_0^1(\lambda)}^2 = 3 \sum_{\ell\in\mathbb{N}} \sum_j 2^{2\ell} |\boldsymbol{v}_{\ell j}|^2.$$

## Appendix C. Best $n$-term rates in higher dimensions

Recall that we consider isotropic weight sequences of the form $\bar{\boldsymbol{\omega}}(a) = \boldsymbol{\omega}(a)^{\otimes M}$, where $a \in (0,\infty)$ determines growth of $\boldsymbol{\omega}(a)$. To obtain worst-case rates for the approximation, we apply Lemmas 2.1 and 2.11

$$\|\boldsymbol{u} - P_{J_n\boldsymbol{u}}\boldsymbol{u}\|_{\ell^2} \le \|P_{J_{n+1}}\bar{\boldsymbol{\omega}}(a)\|_{\ell^2}^{-1} \|\boldsymbol{u}\|_{\ell^1_{\bar{\boldsymbol{\omega}}(a)}} \le \|P_{J_{n+1}}\bar{\boldsymbol{\omega}}(a)\|_{\ell^2}^{-1} \|\bar{\boldsymbol{\omega}}(A)^{-1}\|_{\ell^2_{\bar{\boldsymbol{\omega}}(a)}} \|\boldsymbol{u}\|_{\ell^2_{\bar{\boldsymbol{\omega}}(A)}}$$

and compute an upper bound for the decay rate $\varepsilon(n) := \|P_{J_{n+1}}\bar{\boldsymbol{\omega}}(a)\|_{\ell^2}^{-1}$. Then, for fixed $a$, we choose the parameter $A > a$ as small as possible while ensuring that $\|\bar{\boldsymbol{\omega}}(A)^{-1}\|_{\ell^2_{\bar{\boldsymbol{\omega}}(a)}}$ is finite.

### C.1. Exponential decay (analytic regularity)

First, consider the weight sequence $\boldsymbol{\omega}(a)_j = f_a(j)$ with $f_a(x) := \exp(ax)$ and note that $\bar{\boldsymbol{\omega}}(a)_j = \bar{f}_a(j) := \prod_{m=1}^M f_a(j_m)$. To obtain a worst-case bound for the best $n$-term approximation of sequences $\boldsymbol{u}$ with $\|\boldsymbol{u}\|_{\ell^2_{\bar{\boldsymbol{\omega}}(A)}} = 1$, we seek sets $J_n^*$ of size $n$ that maximise the factor $\|P_{J_n^*}\bar{\boldsymbol{\omega}}(a)\|_{\ell^2}^{-1}$. Finding such a set is equivalent to finding a set that minimises $\|P_{J_n^*}\bar{\boldsymbol{\omega}}(a)\|_{\ell^2}^2$. Since $\bar{f}_a(j) = \exp(a\|j\|_1)$ is monotonic in $\|j\|_1$, we can define for every $R \in \mathbb{N}$ the set $J_R^\circ := \{j \in \mathbb{N}^M : \|j\|_1 \le R\}$, which minimises

$$d(R) := \|P_{J_R^\circ}\bar{\boldsymbol{\omega}}(a)\|_{\ell^2}^2 = \sum_{\|j\|_1\le R} \bar{f}_a(j)^2$$

over all sets $J_R^\circ$ with cardinality bounded by

$$n(R) := |J_R^\circ| = \sum_{\|j\|_1\le R} 1.$$

This means that for every $R \in \mathbb{N}$ we can find a set of size $n(R)$ which results in the error bound $d(R)^{-1/2}$. Solving this relation for $n$ gives an expression for $\varepsilon(n) = d(R(n))^{-1/2}$. Note that this expression is technically only correct if $n = n(R)$ for some $R \in \mathbb{N}$. To obtain an explicit bound that is valid for all values of $n$, we derive monotonic lower and upper bounds $\underline{d}(R) \le d(R)$ and $\overline{n}(R) \ge n(R)$ and define the inverse $\underline{R}(n) := \overline{n}^{-1}(n)$ as well as $\overline{\varepsilon}(n) := \underline{d}(\underline{R}(n))^{-1/2}$. Since these bounds are monotonic, it holds that $\underline{R}(n) \le R(n)$ and thus

$$\varepsilon(n) = d(R(n))^{-1/2} \le \underline{d}(\underline{R}(n))^{-1/2} = \overline{\varepsilon}(n)$$

if $n = n(R)$ for some $R$. Moreover, since $n(R)$ increases monotonically with $R$, we can choose for any $n \in \mathbb{N}$ a value $R \in \mathbb{N}$ such that $n(R) \le n \le n(R+1) \le \overline{n}(R+1)$. Then we can bound

$$\frac{\varepsilon(n)}{\overline{\varepsilon}(n)} \le \frac{\varepsilon(n(R))}{\overline{\varepsilon}(\overline{n}(R+1))} = \left(\frac{\underline{d}(\underline{R}(\overline{n}(R+1)))}{d(R(n(R)))}\right)^{1/2} = \left(\frac{\underline{d}(R+1)}{d(R)}\right)^{1/2} \le \left(\frac{\underline{d}(R+1)}{\underline{d}(R)}\right)^{1/2}.$$

Hence, if $c_\varepsilon := \big(\sup_{R\in\mathbb{N}} \frac{\underline{d}(R+1)}{\underline{d}(R)}\big)^{1/2}$ remains bounded, we obtain for any $n \in \mathbb{N}$ the bound

$$\varepsilon(n) \leq c_\varepsilon \overline{\varepsilon}(n).$$

To find analytic expressions for $\underline{d}$ and $\overline{n}$, we interpret the sums in $d(R)$ and $n(R)$ as Riemann sums

$$d(R) \gtrsim \int_{\substack{\|x\|_1\leq R\\ x\geq 0}} \bar{f}_a(x)^2 \,\mathrm{d}x \qquad \text{and} \qquad n(R) \lesssim \int_{\substack{\|x\|_1\leq R\\ x\geq 0}} 1 \,\mathrm{d}x.$$

To compute the integrals, recall that the volume and surface area of the $M$-dimensional $\ell^1$-ball are given by

$$V_M(R) = 2^M \frac{R^M}{M!} \qquad \text{and} \qquad A_M(R) = 2^M \sqrt{M} \frac{R^(M-1)}{(M-1)!}.$$

Utilising the symmetry of the $\ell^1$-ball, this immediately yields

$$n(R) \lesssim \int_{\substack{\|x\|_1\leq R\\ x\geq 0}} 1 \,\mathrm{d}x = 2^{-M} \int_{\|x\|_1\leq R} 1 \,\mathrm{d}x = 2^{-M} V_M(R) = \frac{R^M}{M!}.$$

To bound $d(R)$, note that Fubini's theorem implies

$$\begin{aligned}
\int_{\substack{\|x\|_1\leq R\\ x\geq 0}} \exp(a\|x\|_1) \,\mathrm{d}x &= \int_0^R \int_{\substack{\|x\|_1=r\\ x\geq 0}} \exp(a\|x\|_1) \,\mathrm{d}x \,\mathrm{d}r = \int_0^R \exp(ar) \int_{\substack{\|x\|_1=r\\ x\geq 0}} 1 \,\mathrm{d}x \,\mathrm{d}r \\
&= \int_0^R \exp(ar) 2^{-M} \int_{\|x\|_1=r} 1 \,\mathrm{d}x \,\mathrm{d}r = \int_0^R \exp(ar) 2^{-M} A_M(r) \,\mathrm{d}r \\
&= \int_0^R \exp(ar) \sqrt{M} \frac{r^{M-1}}{(M-1)!} \,\mathrm{d}r \overset{(*)}{=} \sqrt{M}(-a)^{-M} \left(1 - \exp(aR) \sum_{k=0}^{M-1} \tfrac{(-aR)^k}{k!}\right) \\
&= \sqrt{M} a^{-M} \exp(aR) \left| \exp(-aR) - \sum_{k=0}^{M-1} \tfrac{(-aR)^k}{k!} \right|,
\end{aligned}$$

where the equality $(*)$ follows from the definition of the incomplete gamma function. Note that the last line approaches $\sqrt{M} a^{-M} \exp(aR) \frac{(aR)^{M-1}}{(M-1)!} = \sqrt{M} \frac{R^{M-1}}{a(M-1)!} \exp(aR)$ as $R$ increases. For the sake of simplicity, we hence compute the rates only up to asymptotic equivalence. This yields the bounds

- $\overline{n}(R) = c_n \frac{R^M}{M!}$,
- $\underline{R}(n) = c_R n^{1/M}$ with $c_R := (\frac{M!}{c_n})^{1/M}$ and
- $\underline{d}(R) = c_d \sqrt{M} \frac{R^{M-1}}{2a(M-1)!} \exp(2aR)$.

Moreover, it holds that $c_\varepsilon = \sup_{R\in\mathbb{N}_{\geq 1}} \Big(1 + \frac{1}{R}\Big)^{(M-1)/2} \exp(2a) < \infty$ and, consequently,

$$\varepsilon(n) \leq c_\varepsilon \overline{\varepsilon}(n) = c_\varepsilon c_d n^{-(M-1)/(2M)} \exp(-c_R a n^{1/M}).$$

Finally, observe that $\|\bar{\boldsymbol{\omega}}(A)^{-1}\|_{\ell^2_{\bar{\boldsymbol{\omega}}(a)}} < \infty$ is valid for any $a < A$.

C.2. **Algebraic decay (mixed Sobolev regularity)**

Consider the weight sequence $\boldsymbol{\omega}(a)_j = g_a(j)$ with $g_a(x) := (x+1)^a$ and note that $\bar{\boldsymbol{\omega}}(a)_j = \bar{g}_a(j) := \prod_{m=1}^M g_a(j_m)$. Since $g_a(x) = f_a(\ln(x+1))$, this case can be reduced to the case of exponential decay. The set of indices which minimise the decay rate are $J_R^\circ := \{j \in \mathbb{N}^M : \|\ln(j+1)\|_1 \leq R\}$. Replacing the (Riemann) sums by integrals and performing the change of variables $y := \ln(x+1)$, we obtain

$$d(R) \gtrsim \int_{\substack{\|\ln(x+1)\|_1 \leq R \\ x \geq 0}} \bar{g}_a(x)^2 \,\mathrm{d}x = \int_{\substack{\|y\|_1 \leq R \\ y \geq 0}} \bar{f}_a(y)^2 \exp(\|y\|_1)\,\mathrm{d}y = \int_{\substack{\|y\|_1 \leq R \\ y \geq 0}} \exp((2a+1)\|y\|_1)\,\mathrm{d}y,$$

$$n(R) \lesssim \int_{\substack{\|\ln(x+1)\|_1 \leq R \\ x \geq 0}} 1 \,\mathrm{d}x = \int_{\substack{\|y\|_1 \leq R \\ y \geq 0}} \exp(\|y\|_1)\,\mathrm{d}y.$$

This yields the bounds

- $\overline{n}(R) = c_n\sqrt{M}\frac{R^{M-1}}{(M-1)!}\exp(R)$,
- $R(n) = (M-1)W\Big(c_R n^{1/(M-1)}\Big)$ with $c_R := \frac{1}{M-1}\Big(\frac{(M-1)!}{c_n\sqrt{M}}\Big)^{1/(M-1)}$ and
- $\underline{d}(R) = c_d\sqrt{M}\frac{R^{M-1}}{(2a+1)(M-1)!}\exp((2a+1)R)$.

Moreover, it holds that $c_\varepsilon = \sup_{R\in\mathbb{N}_{\geq 1}}\Big(1+\frac{1}{R}\Big)^{(M-1)/2}\exp(2a+1) < \infty$. To obtain a decay rate, we define $y := c_R n^{1/(M-1)}$ and recall that [34]

$$\ln(y) - \ln\ln(y) \leq W(y) \leq \ln(y) - \tfrac{1}{2}\ln\ln(y) \leq \ln(y)$$

for every $y \geq e$. This implies

$$\begin{aligned}\underline{d}(\underline{R}(n)) &\propto W(y)^{M-1}\exp((2a+1)(M-1)W(y)) \\ &= y^{(2a+1)(M-1)}W(y)^{-2a(M-1)} \\ &\geq y^{(2a+1)(M-1)}\ln(y)^{-2a(M-1)} \\ &\propto n^{2a+1}\ln(n)^{-2a(M-1)},\end{aligned}$$

which yields the bound

$$\varepsilon(n) \lesssim n^{-(a+1/2)}\ln(n)^{a(M-1)}.$$

Finally, observe that $\|\bar{\boldsymbol{\omega}}(A)^{-1}\|_{\ell^2_{\bar{\boldsymbol{\omega}}(a)}} < \infty$ is valid for any $a < A - \frac{1}{2}$.

## Appendix D. The advantage of low ranks for approximation

To illustrate the advantage of this new format, consider approximating the rank-1 function $x \mapsto \exp(x_1 + \ldots + x_M)$ by Legendre polynomials. To solve this approximation problem by means of an ALS-type algorithm, a sequence of microsteps have to be performed that read

$$\operatorname*{minimise}_{\|C\|_{\ell^0_{\boldsymbol{\beta}}} \leq r} \|F - MQC\|^2_{\ell^2}.$$

Here, the vector $F$ and operator $M$ are defined as in (4.5) with $B = \mathrm{vec}(b \otimes \cdots \otimes b)$ given by a vector of tensor product Legendre polynomials $b : [-1,1] \to \mathbb{R}^d$ of degree at most $d-1$. The operator $Q$ maps the core tensor $C$ to the full tensor and corresponds to a choice of basis $Q^\intercal B$ for the least squares problem of the microstep. Note that the weighted sparsity constraint $\|C\|_{\ell^0_{\boldsymbol{\beta}}} \leq r$ is less restrictive the better the sought function $u$ can be expressed in the basis $Q$. It is therefore instructive to compare

the basis $Q$ that is chosen in the $k^{\text{th}}$ microstep of sparse ALS (Algorithm 2) to the minimal basis that is chosen by a classical ALS.

**Sparse ALS.** In the sparse ALS, $Q \in \mathcal{Q}_{R,k}$ is (up to reshaping) an orthogonal matrix where every column is a standard basis vector (cf. Theorem 4.9). This means that the basis functions $\widetilde{B}^{\text{sparse}} := Q^\intercal B$ in this linear least squares problem are of the form

$$\widetilde{B}_j^{\text{sparse}}(x) := B_{\alpha^{(j)}}(x)$$

for some multi-indices $\alpha^{(j)} \in [d]^M$, i.e. that every $\widetilde{B}_j^{\text{sparse}}$ is a product Legendre polynomial of potentially high degree. Since the sought function

$$u(x) = \exp(x_1 + \ldots + x_M) = \exp(x_1) \cdots \exp(x_M) = u_1(x_1) \cdots u_M(x_M)$$

is of rank 1, the best approximation $v = v_1 \otimes \cdots \otimes v_M$ is of rank 1 as well. But the number of terms in this approximation is exponentially large. This must be the case for any approximation with small error, since the approximation error $\|u - v\|_{L^2}$ is equivalent to the approximation error the individual factors

$$\max_k \|u_k - v_k\|_{L^2} \lesssim \|u - v\|_{L^2} \lesssim \|u_1 - v_1\|_{L^2} + \cdots + \|u_M - v_M\|_{L^2} \lesssim \max_k \|u_k - v_k\|_{L^2}.$$

Due to this error equivalence and the symmetry of $u$, every factor $u_k$ must be approximated to the same accuracy by a polynomial $v_k$ of uniform degree $g - 1$. Thus $\widetilde{B}^{\text{sparse}}$ has to be the product basis $\widetilde{B}^{\text{sparse}} = \widetilde{b}^{\otimes(k-1)} \otimes b \otimes \widetilde{b}^{\otimes(M-k)}$, where $\widetilde{b} : [-1, 1] \to \mathbb{R}^g$ ith the vector of Legendre polynomials of degree at most $g - 1$. This means that the basis has to be $(g^{M-1}d)$-dimensional.

**Standard ALS.** Suppose that every component tensor other than the $k^{\text{th}}$ has been updated at least once. Then the current approximation has the form

$$QC = E_1 \otimes \cdots \otimes E_{k-1} \otimes \operatorname{vec}(C) \otimes E_{k+1} \otimes \cdots \otimes E_M,$$

where the vectors $E_\ell$ are the coefficients of one-dimensional Legendre polynomial approximations to the exponential function

$$\widetilde{\exp}_\ell(x) := E_\ell^\intercal b(x)$$

on the interval $[-1, 1]$. The basis function $\widetilde{B}^{\text{dense}} := Q^\intercal B$ for the microstep are hence give by

$$\widetilde{B}_j^{\text{dense}}(x) := b_j(x_k) \prod_{\ell \neq k} \widetilde{\exp}_\ell(x_\ell).$$

**Comparison.** In the preceding two paragraphs we have seen that the basis dimension in the sparse ALS is exponentially larger than in the standard ALS. From a computational point of view, this drastically increases the complexity of the micro steps. From a statistical point of view this also decreases the probability of the RIP. Assuming the approximations $\widetilde{\exp}_\ell \approx \exp$ are sufficiently good, it holds for every $j \geq 4$ that

$$\|\widetilde{\exp}_\ell\|_{L^\infty([-1,1])} \approx \|\exp\|_{L^\infty([-1,1])} = e \leq \sqrt{2j+1} = \|b_j\|_{L^\infty([-1,1])}.$$

This means that $\|\widetilde{B}_j^{\text{dense}}\|_{L^\infty} \leq \|\widetilde{B}_j^{\text{sparse}}\|_{L^\infty}$ (approximately) for the majority of indices $1 \leq j \leq r_{k-1} d r_k$. Since Theorem 1.5 requires $\boldsymbol{\beta}_j \geq \|\widetilde{B}_j^{\text{sparse}}\|_{L^\infty}$ for the sparse ALS and $\boldsymbol{\beta}_j \geq \|\widetilde{B}_j^{\text{dense}}\|_{L^\infty}$ for the standard ALS, the sparsity constraint $\|C\|_{\ell^0_{\boldsymbol{\beta}}} \leq r$ is less restrictive for the standard ALS and the same approximation error can be achieved with a smaller value of $r$. In this special case, rounding would provide a better basis for the sparse approximation, which indicates that reducing the rank may increase the practical performance of the (then less sparse) ALS. In general, however, the basis in the

low-rank representation is not uniquely defined and has to be adapted before performing the sparse approximation.

## References


[1] Ben Adcock and Simone Brugiapaglia. Is Monte Carlo a bad sampling strategy for learning smooth functions in high dimensions? `https://arxiv.org/abs/2208.09045`, 2022.

[2] Ben Adcock, Simone Brugiapaglia, and Clayton G Webster. *Sparse Polynomial Approximation of High-Dimensional Functions*, volume 25 of *Computational Science & Engineering*. Society for Industrial and Applied Mathematics, 2022.

[3] Mazen Ali and Anthony Nouy. Approximation Theory of Tree Tensor Networks: Tensorized Multivariate Functions. `https://arxiv.org/abs/2101.11932`, 2021.

[4] Mazen Ali and Anthony Nouy. Approximation Theory of Tree Tensor Networks: Tensorized Univariate Functions. *Computing*, 58(2):463–544, 2023.

[5] Markus Bachmayr, Albert Cohen, Ronald DeVore, and Giovanni Migliorati. Sparse polynomial approximation of parametric elliptic PDEs. Part II: lognormal coefficients. *ESAIM, Math. Model. Numer. Anal.*, 51(1):341–363, 2016.

[6] Markus Bachmayr, Albert Cohen, and Giovanni Migliorati. Sparse polynomial approximation of parametric elliptic PDEs. Part I: affine coefficients. *ESAIM, Math. Model. Numer. Anal.*, 51(1):321–339, 2016.

[7] Markus Bachmayr, Albert Cohen, and Giovanni Migliorati. Representations of Gaussian Random Fields and Approximation of Elliptic PDEs with Lognormal Coefficients. *J. Fourier Anal. Appl.*, 24(3):621–649, 2017.

[8] Markus Bachmayr, Michael Götte, and Max Pfeffer. Particle number conservation and block structures in matrix product states. *Calcolo*, 59(2):1–47, 2022.

[9] Markus Bachmayr, Anthony Nouy, and Reinhold Schneider. Approximation by tree tensor networks in high dimensions: Sobolev and compositional functions. *Pure Appl. Funct. Anal.*, 8(2):405–428, 2023.

[10] Daniele Bigoni, Allan P. Engsig-Karup, and Youssef M. Marzouk. Spectral Tensor-Train Decomposition. *SIAM J. Sci. Comput.*, 38(4):A2405–A2439, 2016.

[11] Jean-Luc Bouchot, Benjamin Bykowski, Holger Rauhut, and Christoph Schwab. Compressed sensing Petrov-Galerkin approximations for parametric PDEs. In *2015 International Conference on Sampling Theory and Applications (SampTA)*. IEEE, 2015.

[12] Emmanuel J. Candès. The restricted isometry property and its implications for compressed sensing. *C. R. Math.*, 346(9):589–592, 2008.

[13] Gavin C. Cawley and Nicola L. C. Talbot. Fast exact leave-one-out cross-validation of sparse least-squares support vector machines. *Neural Netw.*, 17(10):1467–1475, 2004.

[14] Olivier Chapelle, Vladimir Vapnik, and Yoshua Bengio. Model Selection for Small Sample Regression. *Mach. Learn.*, 48(1/3):9–23, 2002.

[15] Ziang Chen, Jianfeng Lu, Yulong Lu, and Shengxuan Zhou. A Regularity Theory for Static Schrödinger Equations on $\mathbb{R}^d$ in Spectral Barron Spaces. `https://arxiv.org/abs/2201.10072`, 2022.

[16] M. Chevreuil, R. Lebrun, A. Nouy, and P. Rai. A Least-Squares Method for Sparse Low Rank Approximation of Multivariate Functions. *SIAM/ASA J. Uncertain. Quantif.*, 3(1):897–921, 2015.

[17] Albert Cohen and Ronald DeVore. Approximation of high-dimensional parametric PDEs. *Acta Numer.*, 24:1–159, 2015.

[18] Albert Cohen, Ronald Devore, and Christoph Schwab. Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDE's. *Anal. Appl., Singap.*, 09(1):11–47, 2011.

[19] Albert Cohen and Giovanni Migliorati. Optimal weighted least-squares methods. *SMAI J. Comput. Math.*, 3:181–203, 2017.

[20] Albert Cohen and Giovanni Migliorati. Near-optimal approximation methods for elliptic PDEs with log-normal coefficients. *Math. Comput.*, 92(342):1665–1691, 2023.

[21] Ronald A. DeVore. Nonlinear approximation. *Acta Numer.*, 7:51–150, 1998.

[22] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Ann. Stat.*, 32(2):407–499, 2004.

[23] Martin Eigel, Reinhold Schneider, Philipp Trunschke, and Sebastian Wolf. Variational Monte Carlo—bridging concepts of machine learning and high-dimensional partial differential equations. *Adv. Comput. Math.*, 45:2503–2532, 2019.

[24] Michael Götte, Reinhold Schneider, and Philipp Trunschke. A block-sparse Tensor Train Format for sample-efficient high-dimensional Polynomial Regression. `https://arxiv.org/abs/2104.14255`, 2021.

[25] Lars Grasedyck and Sebastian Krämer. Stable ALS approximation in the TT-format for rank-adaptive tensor completion. *Numer. Math.*, 143(4):855–904, 2019.

[26] Erwan Grelier, Anthony Nouy, and Mathilde Chevreuil. Learning with tree-based tensor formats. `https://arxiv.org/abs/1811.04455`, 2018.

[27] Michael Griebel and Helmut Harbrecht. On the constrction of sparse tensor product spaces. *Math. Comput.*, 82(282):975–994, 2013.

[28] Michael Griebel and Helmut Harbrecht. Singular value decomposition versus sparse grids: refined complexity estimates. *IMA J. Numer. Anal.*, 39(4):1652–1671, 2018.

[29] Michael Griebel and Helmut Harbrecht. Analysis of tensor approximation schemes for continuous functions. `https://arxiv.org/abs/1903.04234`, 2021.

[30] Michael Griebel, Helmut Harbrecht, and Reinhold Schneider. Low-rank approximation of continuous functions in Sobolev spaces with dominating mixed smoothness. `https://arxiv.org/abs/2203.04100`, 2022.

[31] Markus Hansen and Winfried Sickel. Best m-term aproximation and tensor product of Sobolev and Besov spaces-the case of non-compact embeddings. *East J. Approx.*, 16:345–388, 2010.

[32] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020.

[33] Sebastian Holtz, Thorsten Rohwedder, and Reinhold Schneider. The Alternating Linear Scheme for Tensor Optimization in the Tensor Train Format. *SIAM J. Sci. Comput.*, 34(2):A683–A713, 2012.

[34] Abdolhossein Hoorfar and Mehdi Hassani. Inequalities on the Lambert function and hyperpower function. *JIPAM, J. Inequal. Pure Appl. Math.*, 9(2): article no. 51 (5 pages), 2008.

[35] J. D. Hunter. Matplotlib: A 2D graphics environment. *Comput. Sci. & Eng.*, 9(3):90–95, 2007.

[36] Arieh Iserles. A fast and simple algorithm for the computation of Legendre coefficients. *Numer. Math.*, 117(3):529–553, 2010.

[37] Christopher Leisner. Nonlinear Wavelet Approximation in Anisotropic Besov Spaces. *Indiana Univ. Math. J.*, 52(2):437–455, 2003.

[38] Lingjie Li, Wenjian Yu, and Kim Batselier. Faster tensor train decomposition for sparse data. *J. Comput. Appl. Math.*, 405: article no. 113972, 2022.

[39] Gabriel J Lord, Catherine E Powell, and Tony Shardlow. *An introduction to computational stochastic PDEs*. Cambridge Texts in Applied Mathematics. Cambridge University Press, 2014.

[40] Bertrand Michel and Anthony Nouy. Learning with tree tensor networks: Complexity estimates and model selection. *Bernoulli*, 28(2):910–936, 2022.

[41] A. Nouy. *Low-Rank Methods for High-Dimensional Approximation and Model Order Reduction*, chapter 4. Society for Industrial and Applied Mathematics, 2017.

[42] Anthony Nouy and Erwan Grelier. tensap, 2020. `https://doi.org/10.5281/zenodo.3894378`.

[43] Ivan V. Oseledets. Tensor-Train Decomposition. *SIAM J. Sci. Comput.*, 33(5):2295–2317, 2011.

[44] Earl David Rainville. *Special functions*. The Macmillan Company, 1960.

[45] Holger Rauhut and Christoph Schwab. Compressive sensing Petrov-Galerkin approximation of high-dimensional parametric operator equations. *Math. Comput.*, 86(304):661–700, 2016.

[46] Holger Rauhut and Rachel Ward. Interpolation via weighted $\ell^1$ minimization. *Appl. Comput. Harmon. Anal.*, 40(2):321–351, 2016.

[47] Abel Sancarlos, Victor Champaney, Jean-Louis Duval, Elias Cueto, and Francisco Chinesta. Pgd-based advanced nonlinear multiparametric regressions for constructing metamodels at the scarce-data limit. `https://arxiv.org/abs/2103.05358`, 2021.

[48] Fadil Santosa and William W. Symes. Linear Inversion of Band-Limited Reflection Seismograms. *SIAM J. Sci. Stat. Comput.*, 7(4):1307–1330, 1986.

[49] R. Schneider and A. Uschmajew. Approximation rates for the hierarchical tensor format in periodic Sobolev spaces. *J. Complexity*, 30(2):56–71, 2014.

[50] Christoph Schwab and Claude Jeffrey Gittelson. Sparse tensor discretizations of high-dimensional parametric and stochastic PDEs. *Acta Numer.*, 20:291–467, 2011.

[51] Sukhwinder Singh, Robert N. C. Pfeifer, and Guifré Vidal. Tensor network decompositions in the presence of a global symmetry. *Phys. Rev. A*, 82(5): article no. 050301(R), 2010.

[52] Radu Alexandru Todor and Christoph Schwab. Convergence rates for sparse chaos approximations of elliptic problems with stochastic coefficients. *IMA J. Numer. Anal.*, 27(2):232–261, 2007.

[53] Philipp Trunschke. Convergence bounds for nonlinear least squares and applications to tensor recovery. `https://arxiv.org/abs/2108.05237`, 2021.

[54] Philipp Trunschke. Convergence bounds for nonlinear least squares for tensor recovery. `https://arxiv.org/abs/2208.10954`, 2022.

[55] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods*, 17:261–272, 2020.

[56] Herbert F Wang and Mary P Anderson. *Introduction to groundwater modeling: finite difference and finite element methods*. Academic Press Inc., 1995.

[57] Wenqi Wang, Vaneet Aggarwal, and Shuchin Aeron. Tensor Train Neighborhood Preserving Embedding. *IEEE Trans. Signal Process.*, 66(10):2724–2732, 2018.